Towards Less Supervision for Scalable Recognition of Daily Activities

A dissertation submitted to
TECHNISCHE UNIVERSITÄT DARMSTADT
Fachbereich Informatik

for the degree of
Doktor-Ingenieur (Dr.-Ing.)

presented by

MAJA STIKIC
Dipl.-Ing.

born 6th of May, 1978
in Sarajevo, Bosnia and Herzegovina

Prof. Dr. Bernt Schiele, examiner
Prof. Dr. Thad Starner, co-examiner

Date of Submission: 29th of April, 2009
Date of Defense: 15th of June, 2009

Darmstadt, 2010
D17

Abstract

This thesis is concerned with scalable recognition of human activities in real-world settings. Research towards this aim has addressed the automated detection of Activities of Daily Living, such as personal hygiene, eating, meal preparation, or housekeeping, as a particularly fruitful endeavor for elderly health care. The focus of this thesis lies on two challenges within these efforts: characterization of daily activities in sensor readings, and practical methods to label these data. We address these challenges by investigating several research directions for unobtrusive activity recognition that require only a limited number of sensors and minimal annotation overhead.

We utilize a multi-sensor approach to characterize two important aspects of activities. We use wearable acceleration sensors to infer characteristic body movements and RFID tags in combination with RFID readers to recognize object usage during execution of activities. The benefit of the proposed approach is that it is able to attain high recognition performance even when the number of sensors is significantly decreased to a single wrist-worn sensor and just a few tagged objects. This is achieved by augmenting the learning process with additional information from complementary sensors.

We also explore the combination of two types of sensors, namely accelerometers for body-motion and infra-red sensors for detecting indoor location where the activities are performed. The goal of this study is to investigate the applicability of two different techniques to significantly reduce the need for labeled training data. The first technique combines small amounts of labeled activity data with easily obtainable unlabeled data in a semi-supervised learning process. The second technique aims at focusing labeling efforts on the most profitable instances by utilizing active learning. The experimental results indicate that we can achieve comparable and sometimes even better performance than the fully supervised approaches.

In order to further enhance the applicability of activity recognition in real-world settings, we propose a novel multi-instance learning method that is able to learn from sparsely labeled data. Instead of requiring labels for each individual training sample, we group sensor data into bags-of-activities and provide the labels only on the bag level. We propose several novel algorithmic extensions of multi-instance learning that support new labeling scenarios allowing less constrained ways of annotating activity data. We systematically analyze the trade-off between the labeling efforts and recognition performance.

Lastly, we introduce several graph-based label propagation strategies for enabling long-term activity recordings without the need for detailed continuous activity annotations. We propose two different ways of combining multiple graphs based on data similarity in feature space and time. We carry out a comparative evaluation of this approach and the multi-instance learning approach. We show that the graph-based approach outperforms multi-instance learning.

Overall, this thesis demonstrates the feasibility of using unlabeled data for learning more expressive activity classifiers and the potential of multi-sensor approaches to facilitate scalable activity recognition.

Zusammenfassung

Diese Arbeit beschäftigt sich mit der skalierbaren Erkennung menschlicher Aktivitäten in anwendungsnahen Situationen. Forschung, ausgerichtet auf die Erkennung von Aktivitäten des täglichen Lebens (ATL), wie zum Beispiel eigene Körperpflege, Nahrungsaufnahme, Essenszubereitung und Führung des Haushalts, stellt sich als interessante Fragestellung im Rahmen der Altenpflege dar. Diese Arbeit konzentriert sich auf zwei Herausforderungen: Die Charakterisierung täglicher Aktivitäten in Sensordaten und eine einfach durchzuführende Annotation dieser Daten. Wir begegnen diesen Herausforderungen, indem wir verschiedene Techniken zur unaufdringlichen Erkennung von Aktivitäten entwickeln, welche zudem nur eine geringe Anzahl von Sensoren und einen minimalen Annotationsaufwand erfordern.

Für die Charakterisierung von Aktivitäten verwenden wir einen Multi-Sensor Ansatz. Wir setzen tragbare Beschleunigungssensoren ein, um Rückschlüsse auf charakteristische Körperbewegungen zu ziehen, sowie RFID Tags in Kombination mit RFID Lesegeräten, um die Verwendung verschiedener Objekte während der Ausführung einer Aktivität zu detektieren. Der Vorteil dieser Vorgehensweise liegt darin, dass sie selbst dann hohe Erkennungsraten aufweist, wenn die Anzahl der getragenen Sensoren und markierten Objekte signifikant reduziert wird, so dass z.B. ein einziger Beschleunigungssensor am Handgelenk und wenige Objekte genügen. Möglich wird dies durch die Kombination der Informationen der beiden komplementären Sensortypen.

Wir untersuchen weiterhin eine andere Kombination zweier Sensortypen, und zwar Beschleunigungssensoren für Körperbewegungen und Infrarot-Sensoren für die Ortsbestimmung des Nutzers innerhalb eines Gebäudes. Das Ziel dieser Untersuchung ist die Analyse zweier Techniken zur Reduzierung annotierter Trainingsdaten. Die erste Technik kombiniert eine geringe Menge an annotierten Daten mit leicht erlangbaren unannotierten Daten in einem Halb-Überwachten Lernprozess (eng. Semi-supervised Learning). Die zweite Technik zielt darauf ab, den Aufwand für die Annotationen auf die lohnensten Datenpunkte zu konzentrieren. Dies geschieht mit einer Methode namens Aktives Lernen (eng. Active Learning). Die experimentellen Ergebnisse deuten darauf hin, dass wir im Vergleich mit anderen, vollständig überwachten Ansätzen mit unseren Techniken eine vergleichbare und teils sogar bessere Performanz erreichen.

Um die Einsetzbarkeit von Aktivitätserkennung in anwendungsnahen Situationen weiter zu erhöhen, schlagen wir einen neuen Multi-Instanz Lernansatz vor, welcher in der Lage ist von spärlich annotierten Daten zu lernen. Statt Annotationen für jedes Trainingsbeispiel vorauszusetzen, gruppieren wir die Sensordaten in Gruppen von Aktivitäten (eng. bags-of-activities) und vergeben Annotationen nur auf der Ebene dieser Gruppen. Wir schlagen ferner mehrere Erweiterungen des Multi-Instanz Ansatzes vor, welche neuartige Annotations-Methoden mit weniger Beschränkungen als herkömmliche Verfahren erlauben. Ausserdem analysieren wir die notwendige Abwägung zwischen Annotationsaufwand und Erkennungsleistung.

Im letzten Teil der Arbeit führen wir mehrere graphen-basierte Strategien zur Propagierung von Annotationen ein. Das Ziel hierbei ist die Ermöglichung von langfristigen

Aktivitätsaufnahmen ohne die Erfordernis von detaillierten Annotationen der ausgeführten Aktivitäten. Wir schlagen zwei verschiedene Techniken vor, mehrere Graphen anhand von Ähnlichkeiten im Merkmalsraum und in der Zeit zu kombinieren. Wir führen eine vergleichende Studie unserer Technik und dem Multi-Instanz Ansatz durch und zeigen, dass der graphen-basierte Ansatz dem Multi-Instanz Ansatz überlegen ist.

Die wichtigsten Beiträge dieser Arbeit sind zum einen der Nachweis, dass nicht-annotierte Sensordaten für das Lernen von ausdrucksstarken Klassifikatoren genutzt werden können, sowie das Aufzeigen, dass Multi-Sensor Ansätze grosses Potential im Bezug auf skalierbare Aktivitätserkennung besitzen.

Acknowledgments

This thesis would not have been possible without the support of my supervisor Professor Bernt Schiele. I would like to sincerely thank him for giving me the opportunity to become a part of the MIS group. At our very first meeting he managed to show me what research should really be about and he has been a great source of inspiration ever since. He always devoted time for our research meetings and provided countless advice and guidance during my research journey. Special thanks go to Professor Thad Starner for his interest in my work and for agreeing to be the co-examiner of this thesis.

I consider myself very lucky for having the privilege of working with an amazing group of young researchers at TU Darmstadt: Micha Andriluka, Ulf Blanke, Victoria Carlsson, Gyuri Dorkó, Sandra Ebert, Mario Fritz, Nikodem Majer, Konrad Schindler, Paul Schnitzspan, Edgar Seemann, Michael Stark, Ulrich Steinhoff, Stefan Walk, Christian Wojek, and Andreas Zinnen. It was a great pleasure to work with you guys! We had a lot of fun and stimulating discussions at our retreats as well as at our many lunch, coffee, dinner, bakery, ice-cream, and other breaks. Especially, I would like to thank my colleague Tâm Huỳnh for being an always calm, unbelievably patient, and caring coworker. He helped me getting started in the beginning of my research and provided plenty of input for our joint publication presented in Chapter 4. Kristof Van Laerhoven provided endless hardware support during the Housekeeping dataset recordings (Chapter 3). I had many fruitful discussions with Diane Larlus about graph-based semi-supervised learning described in Chapter 7. She impressed me with her ability of keeping a sharp eye on the details. Special thanks go to our secretary Ursula Paeckel who made my life much easier by efficiently dealing with administration and paperwork for me.

This work has been supported in part by Fraunhofer's Doktorandinnenprogram. I would like to express my gratitude to the colleagues at Fraunhofer IPSI: Dr. Dr. Norbert Streitz, Carsten Magerkurth, Richard Etter, Carsten Röcker, and Sylvia Lang for their encouragement during the closure of the institute. Many thanks go to Thorsten Prante and Barbara Lutes who convinced me to make a big step of moving to Germany. That had a great impact on my life and I would lack an invaluable experience if I did not follow their suggestion. I would also like to thank Professor Dieter Fellner for agreeing on my official transfer to Fraunhofer IGD and allowing me to focus on my PhD work.

Special thanks go to Intel Research for making their iBracelet available for our experiments as well as to Beth Logan, Jennifer Healey, Emmanuel Munguia Tapia, and Stephen Intille for their assistance and providing access to the PLCouple1 dataset.

My colleagues from Mihajlo Pupin Institute in Belgrade: Zlatko, Meri, Dule, Nebojša, Goca, Aco, and Dragan have made my first work experience very pleasant making my decision to leave the institute much harder.

Last, but not least, I would like to thank my family and friends for their unconditional love and support that gave me the strength to follow my dreams even in the toughest times. I dedicate this thesis to my grandfather Nikola Garača who is missing in my and my family's life since May 2008.

Contents

1

Introduction

Nowadays, mobile devices, consumer electronics, and even household appliances are equipped with microprocessors, software, and networking capabilities allowing Mark Weiser's vision of *ubiquitous computing* [Weiser 1991] to be within reach. The objective of ubiquitous computing is to work unobtrusively in the background and serve people in their everyday lives. In order to achieve that goal, computers should be able to implicitly perceive everyday situations and adapt their behavior accordingly without explicit user interaction. This resulted in an emerging research field called *context-aware computing* [Schilit *et al.* 1994]. As defined by [Dey 2000], context is any information that can be used to characterize the situation of an entity. Location, identity, time, and activity are important context types for characterizing the situation of a particular entity. As user activity is a valuable piece of context information, research on *activity recognition* has attracted increasing attention in recent years. Typical approaches combine sensors with machine learning techniques to recognize the user's activity.

Scalable and unobtrusive recognition of human activities offers a number of opportunities for novel applications that would empower and increase the quality of everyday life. A range of compelling applications in industrial domains arises for supporting workers in their everyday tasks [Lukowicz *et al.* 2007]. In the educational domain, new learning tools are being explored ranging from augmenting lecture environments with automated tools for capturing different streams of classroom activities [Abowd *et al.* 1996], over enabling casual learning of foreign languages [Beaudin *et al.* 2007] or playing piano [Huang *et al.* 2008] throughout a day to helping practicing American Sign Language skills in young children [Brashear *et al.* 2006]. Activity recognition is also appealing for the next generation of applications in the sport context (e.g. [Kunze *et al.* 2006, Chang *et al.* 2007]) to provide trainers with additional information about the progress of the athletes. An important class of applications based on activity recognition is in the medical and health related domains such as promoting an active healthy lifestyle [Consolvo *et al.* 2008b], detecting severe medical conditions and potential threats such as falling [Jafari *et al.* 2007], or supporting hospital staff in their daily routines [Bardram and Christensen 2007].

The main motivation for the work presented in this thesis is *automatic* health assessment. As the elderly population is rapidly growing, new health care challenges are emerg-

ing. A strong desire by aging individuals to remain independent in their homes as long as possible has initiated innovative technological solutions for *assisted living* [AAL]. In order to improve the quality of life of the senior citizens, the new solutions must offer the required medical and home care assistance. This is challenging because elderly people as the target user group might not be familiar with modern information technologies, which could cause acceptance issues. However, as computers are becoming truly ubiquitous, they are starting to be accepted even by elderly. In the future, user acceptance of such systems will be even higher having in mind that the next generations will already be accustomed to computing technology.

Different kinds of cognitive impairments are a frequently occurring health problem among elderly and their early diagnosis is highly important. The common first symptom of age-related diseases is a change in the person's daily behavior. For that purpose, two specific sets of activities have been defined [Katz 1983]. The first set describes the activities that are crucial for self-care and indicate the functional status of a person, which are called *Activities of Daily Living (ADL)*: bathing, dressing, toileting, transferring, continence, and feeding. The second set includes activities that enable the individual to live independently within a community through interaction with the physical and social environment, and are called *Instrumental Activities of Daily Living (IADL)*: using telephone, shopping, food preparation, housekeeping, doing laundry, transportation, taking medications, and handling finances. These two sets of activities are used to evaluate what type of necessary services an individual may need for further semi-independent living at home.

The ADL/IADL assessment is done based on the *scales of independence* (see Appendix A). Today, the assessment is typically done manually by trained caregivers and case managers through interviews and specialized questionnaires. The manual assessment is time consuming and error prone due to numerous reasons. First, the targeted sets of activities include a large set of different activities and it is hard for an elderly person to remember when and how often the activities have been performed in between the two assessments. Second, the symptoms are often denied by a person for a long time [Morris *et al.* 2003, Morris *et al.* 2005]. Third, a person might not say the truth during the interview because of the fear of being transferred to a nursing home [Wilson *et al.* 2005]. Thus, the automatic assessment might significantly ease the work of the caregivers and case managers and improve the accuracy of the assessment itself.

The goal of this thesis is to enable *scalable* long-term monitoring and recognition of daily activities that would support automatic ADL/IADL assessment. For that purpose, we aim to move beyond controlled laboratory experiments and explore the applicability of activity recognition in real-world settings. We concentrate on increasing user acceptance by employing unobtrusive techniques that require only a *limited number of sensors* and *minimal user involvement* in the activity recognition process.

In the following, we present the main challenges imposed on long-term monitoring and automatic recognition of daily activities (Section 1.1) and give an overview of the thesis' contributions (Section 1.2) in addressing these challenges. We conclude the chapter with the outline of the thesis (Section 1.3).

1.1 Challenges

In this section we state the main real-world challenges that this thesis aims to address to enable scalable recognition of daily activities.

1.1.1 Characterization of Daily Activities

Typical daily activities such as ADLs and IADLs are rather complex activities that are composed of different sub-activities whose order might vary. As we will see in Chapter 4, they exhibit high intra-class variability, since there is a large diversity of ways the same activity is being performed by different persons. People perform activities differently depending on location, surrounding people, time, and handedness (i.e. whether they are left- or right-handed). Sometimes even the interpretation of activities varies among people. All this makes the general characterization of daily activities across people very challenging. There is even variability in the multiple executions of the same activity by a single person.

Average *duration* and *occurrence frequency* of daily activities vary greatly among different activities. The activities can last between a few minutes only (e.g. taking medications) up to a couple of hours (e.g. shopping or food preparation). While some activities occur on a daily basis (e.g. eating), others are less frequent (e.g. doing laundry).

An additional difficulty is that activities can be performed in an *overlapping* or *interleaving* manner. When two or more activities are being performed at the same time, e.g. eating a snack while watching TV, they are defined as overlapping activities. Activities are interleaved when the next activity starts before the previous activity is completed. The interrupted activity is usually continued at a later time. An example of interleaving activities is cleaning the house when different activities such as dishwashing, folding laundry, and putting things away might all be interleaved.

Generally, there are three important characteristics of daily activities:

1. *Motion* or *posture* of the body during the execution of activities
2. Usage of different *objects* in various activities
3. *Location* where the activities are being performed

However, none of these three characteristics can be used for a unique representation of an activity. For example, multiple activities might require similar hand movements (e.g. brooming and vacuuming). Also, the same object might be used in different activities (e.g. a cup is used for drinking, but it also has to be cleaned afterwards). There is typically a one-to-one mapping between daily activities and location where they are performed. For instance, dishwashing is always performed in front of the sink or dishwasher, hygiene related activities typically occur in the bathroom, etc. However, there is still lack of location specificity for many activities. One might eat at home, at the office desk, or in

the restaurant. Furthermore, a restaurant is a workplace for the restaurant personnel. Yet, for another person, a restaurant might be a good place for holding a business meeting.

Thus, a promising approach to recognize daily activities is to combine different sensor modalities for characterizing different properties of activities. However, the more sensors are being used the lower is the user acceptance of a system.

1.1.2 Annotation Issue

There has been a clear tendency in the field of activity recognition to move towards real-world settings due to several reasons. First, in controlled laboratory experiments, bias can easily be introduced because the subjects are aware of the presence of the activity recognition system. Second, by restricting the subjects to the predefined sets of scripted activities, the problem might be oversimplified. Third, the experiments include typically only a few subjects, often the researchers themselves who developed the system. They can unconsciously perform the activities in a manner favoring the recognition system. Thus, long-term activity recordings would enable more representative data that better reflect realistic application scenarios.

However, capturing *ground truth* in real-world settings proves to be a non-trivial task. As most existing approaches to activity recognition rely on *supervised* machine learning methods, they require substantial amounts of *labeled* activity data for *training* a classifier. Obtaining accurate and detailed *annotations* of activities is a great challenge for these approaches limiting their applicability and scalability to large amounts of activities and users.

Labeling data for activity recognition systems is a challenging problem for at least two reasons. First, most of the annotation techniques are time-consuming and error-prone. And second, to obtain reliable annotations one has essentially two choices. Either one may rely on invasive sensors such as cameras and microphones which are often not acceptable due to privacy reasons and the time required for annotating data that way. Or, in long-term realistic recordings one typically has to put a labeling burden on a user which is tedious or disrupting for users in particular when detailed annotations are needed. On the other hand, user-annotated data might be erroneous and the mere fact that the user is annotating the activity might change the activity itself.

In many practical problems, data labeling is expensive, but a large amount of unlabeled data can be easily obtained. For this reason, *semi-supervised learning* has been proposed as an alternative in machine learning research. The ultimate goal of semi-supervised learning is to learn from both labeled and unlabeled data. Since many human activities of interest are performed on a daily basis, it is relatively easy to produce large quantities of unlabeled activity data. Thus, semi-supervised learning naturally lends itself to activity recognition. Moreover, it would be highly beneficial to reduce the level of user disruptions by annotating only the most profitable instances. In that context, *active learning* becomes an important alternative in activity recognition.

1.1.3 Other Challenges

In the following, we outline two additional activity recognition challenges that are tackled in this thesis.

Hardware for Long-Term Recordings. Logging of activity data for extended periods of time poses necessary hardware design requirements including robustness, low power consumption, and wearability for on-body sensing. This thesis does not focus on hardware design, but we tackle this challenge by evaluating two sensor platforms (namely, *iBracelet* [Fishkin *et al.* 2005] and *Porcupine* [Van Laerhoven and Aronsen 2007]) and by reporting deployment issues in Chapter 4.

Standardized Evaluation Procedure. Activity recognition is still a relatively new research topic. An important challenge for current research in this field is the lack of a standardized evaluation procedure that would enable a unified way of comparison of different approaches. Typically, an approach is application dependent and is evaluated on a single dataset, recorded for that particular experiment. There have been attempts to build benchmark datasets [Junker *et al.* 2004] and to introduce a standardized evaluation procedure [Minnen *et al.* 2006b, Ward *et al.* 2006] in activity recognition. However, there are only a few datasets publicly available for joint comparison of different algorithms. We strongly believe that the activity recognition community should move forward to more challenging and realistic circumstances. We tackle this challenge by making our dataset publicly available [Stikic and Van Laerhoven 2007] and by evaluating our algorithms on two other public datasets (*PLCouple1* [Logan *et al.* 2007] and *TU Darmstadt* [Huỳnh *et al.* 2008] - see Chapter 3) to avoid potential bias to a particular dataset only.

1.2 Contributions

In the following we summarize the research questions that this thesis aims to answer. In the first part of the thesis (Chapter 4 and Chapter 5) we employ a multi-sensor approach based on the combination of body-worn sensors and sensors placed in the environment to address the following questions:

- Can we reduce the *number of sensors* needed for accurate activity recognition by fusing two different sensor modalities?
- Can we reduce the *level of supervision* in activity recognition by using complementary sensors?

In the second part of the thesis (Chapter 6 and Chapter 7) we aim to further reduce the level of supervision in activity recognition by using only on-body sensing and exploring different label propagation strategies in order to answer the following question:

- Can we recognize activities from *sparsely* labeled data?

Next, we describe the contributions in more details:

First, we investigate the ***combination of RFID and accelerometer sensing*** for activity recognition. The RFID part of the system infers relevant interactions with objects and the accelerometer part of the system recognizes characteristic hand movements. We experimentally show that recognition accuracy can be significantly improved by fusing these two different types of sensors. At the same time the ***number of required sensors is decreased*** by limiting the hardware to a single wrist-worn device and tagging only the key objects. We analyze different acceleration features and algorithms, and we suggest the best tags' placements and the key objects to be tagged for each activity.

Second, we further explore a ***multi-sensor approach*** by using two complementary sensors, namely accelerometers for body-motion and infra-red motion detectors for inferring indoor location on a room level. We systematically analyze two different techniques to ***significantly reduce the required amount of labeled training data***. The first technique is based on semi-supervised learning and uses self-training and co-training. The second technique is inspired by active learning. In this approach the system actively asks which data the user should label. With both techniques, the required amount of training data can be reduced significantly while obtaining similar and sometimes even better performance than standard supervised techniques.

Third, we introduce a novel method for ***activity recognition from sparsely labeled data***. The method is based on ***multi-instance learning*** allowing to significantly reduce the required level of supervision. In particular we propose several novel extensions of multi-instance learning to support different annotation strategies. The validity of the approach is demonstrated on two public datasets for three different labeling scenarios. The experimental results show that this approach can obtain high recognition performance even though only coarse-grained annotations are provided.

Fourth, we introduce a ***graph-based semi-supervised*** approach for scalable recognition of daily activities. The method propagates information through a graph that contains both labeled and unlabeled data. We propose two different ways of ***combining multiple graphs based on feature similarity and time***. We evaluate both the quality of the label propagation process itself and the performance of classifiers trained on the propagated labels. Experimental results indicate that this approach outperforms the previously introduced multi-instance learning approach and in some cases even outperforms fully supervised approaches.

Parts of this thesis have been published in refereed conference and workshop papers. The issues related to the initial dataset recordings (Chapter 4) are reported in [Stikic and Van Laerhoven 2007]. The activity recognition approach based on the combination of RFID and accelerometer sensing (Chapter 4) has been published in [Stikic *et al.* 2008a]. The multi-sensor approach for reducing the level of supervision (Chapter 5) is presented in [Stikic *et al.* 2008b]. The multi-instance approach for activity recognition from sparsely labeled data (Chapter 6) has been published in [Stikic and Schiele 2009]. The graph-based semi-supervised approach for further reducing the level of supervision in activity recognition (Chapter 7) is introduced in [Stikic *et al.* 2009].

1.3 Thesis Outline

The structure of the remainder of this thesis is as follows:

Chapter 2 – *Related Work* reviews related work relevant for this thesis. We begin by a short historical overview of the activity recognition field and then review different applications for health care and elderly care, multi-sensor approaches for activity recognition, annotation techniques, and machine learning algorithms that have been proposed in the literature.

Chapter 3 – *Methodology* describes the three public datasets and several classifiers used in the experiments, as well as our evaluation procedure used in the following chapters.

Chapter 4 – *Combination of RFID and Accelerometer Sensing* presents a scalable multi-sensor approach for activity recognition that integrates the object usage based approach and the body motion based approach in a unified learning scheme. We motivate the proposed combination of sensors by showing the shortcomings of both sensor modalities separately and by proposing an integrated approach that uses a minimal number of sensors and still compensates for the shortcomings of both approaches.

Chapter 5 – *Towards Less Supervision Based on Complementary Sensors* suggests different strategies for reducing the level of supervision in activity recognition by using two complementary sensor modalities. We present a comparative evaluation of four different algorithms on a publicly available dataset and analyze the performance of the algorithms for different amounts of labeled training data. We also compare recognition performance of our algorithms to a fully supervised approach.

Chapter 6 – *Activity Recognition from Sparsely Labeled Data* takes a next step towards further reducing the level of supervision in more realistic settings. We explore multi-instance learning from sparsely labeled activity data in the context of three different annotation scenarios. The trade-off between labeling efforts and activity recognition performance is analyzed and discussed.

Chapter 7 – *Multi-Graph Label Propagation for Activity Recognition* investigates several label propagation strategies for long-term activity recordings enabling scalable recognition of daily activities. We propose a graph-based semi-supervised approach that propagates the few provided labels to the neighboring data points based on time and feature similarity. The approach is compared with the multi-instance learning approach from Chapter 6.

Chapter 8 – *Conclusion and Outlook* summarizes the work presented in this thesis by reviewing the main conclusions and giving an outlook on future work.

2

Related Work

In this chapter we review the state-of-the-art in activity recognition, with special emphasis on issues and challenges which are of interest for this thesis, namely applications in the medical domain which are the main motivation for our work, multi-sensor approaches, annotation techniques, and machine learning methods for activity recognition.

The chapter is organized as follows. In Section 2.1 we give a short historical overview how the field of activity recognition has evolved over the years. Section 2.2 discusses different applications for health care and elderly care that make use of activity recognition. In Section 2.3 we outline different types of sensors used for activity recognition with focus on the multi-sensor approaches. Section 2.4 presents and identifies the shortcomings of typical annotation techniques used during activity recordings. Finally, Section 2.5 introduces the machine learning algorithms that have been applied for activity recognition.

2.1 Activity Recognition Overview

With the constant advances in hardware technology in terms of size, cost, power consumption and processing power, the research focus has shifted away from the traditional desktop computing towards new paradigms of ubiquitous [Weiser 1991] and wearable [Starner 1999] computing. Activity recognition has emerged as an important part of these new research efforts due to its usefulness for context-aware systems [Schilit *et al.* 1994, Abowd *et al.* 1997, Schmidt 2002]. In the late 1990's the preliminary results on the small datasets of relatively simple physical activities such as walking, running, sitting, walking upstairs, walking downstairs, standing, and jumping have been reported for the first time [Farringdon *et al.* 1999, Van Laerhoven and Cakmakci 2000, Randell and Muller 2000, Mäntyjärvi *et al.* 2001]. Based on the promising results of these first preliminary studies, researchers have started working with more robust sensing platforms and more elaborated datasets on the same low-level physical activities (e.g. [Lee and Mase 2002, Van Laerhoven *et al.* 2003, Kern *et al.* 2003, Bao and Intille 2004, Maurer *et al.* 2006, Ravi *et al.* 2005, Lester *et al.* 2006, Pärkkä *et al.* 2006]) achieving impressive results in recent years.

Recognition of more specialized activities fits into different industrial application domains. That has motivated research towards more challenging scenarios such as recognition of office activities (e.g. phone conversation, giving a presentation, or face-to-face conversation) [Oliver *et al.* 2002, Oliver and Horvitz 2005], or assembly tasks such as wood workshop activities [Ward *et al.* 2005, Lukowicz *et al.* 2004], furniture assembly [Antifakos *et al.* 2002] or car, bicycle and aircraft maintenance tasks [Stiefmeier *et al.* 2006, Lukowicz *et al.* 2007, Zinnen *et al.* 2007, Ogris *et al.* 2008]. Moreover, there have been efforts to recognize soldier activities [Minnen *et al.* 2007] in order to augment post-patrol reports.

In sports, different types of activities such as martial arts movements in Tai Chi [Kunze *et al.* 2006] and Kung Fu [Chambers *et al.* 2002], dumbbell exercises [Minnen *et al.* 2006a], free-weight exercises [Chang *et al.* 2007], juggling [Huỳnh and Schiele 2006b], and physical gymnasium activities [Tapia *et al.* 2007a, Ermes *et al.* 2008] have started to be explored.

As already stated in Chapter 1, automated recognition of Activities of Daily Living (ADL) and Instrumental Activities of Daily Living (IADL) gained increasing attention in the activity recognition community (e.g. [Tapia *et al.* 2004, Philipose *et al.* 2004, Bao and Intille 2004, Wyatt *et al.* 2005, Duong *et al.* 2005]). There have been successful attempts to recognize various selected instances of different ADL/IADL classes such as *hand washing* [Hoey *et al.* 2007], *eating* [Gao *et al.* 2004, Amft *et al.* 2007], *cooking* [Tran and Mynatt 2002], *taking medications* [Wan 1999], or *bathroom activities* [Chen *et al.* 2005]. However, there have been no attempts to recognize in-depth the housekeeping activities, which are an important and often occurring IADL class and are the focus of our experiments in Chapter 4. Although this kind of activities is typically not done by the elderly people in the later stages of a disease, its assessment can help significantly in the early detection of symptoms of different age-related diseases.

A majority of the activity recognition studies have been conducted in constrained laboratory settings. In order to enable more realistic non-laboratory studies, a new pioneering research initiative of living-labs has arisen, such as Neural Network House [Mozer 1998], Aware Home [Abowd *et al.* 2000], PlaceLab [Intille *et al.* 2006], and inHaus [Meyer *et al.* 2008]. These instrumented facilities represent valuable testbeds allowing for user and technology studies in settings more natural than a typical laboratory. In such a recent study of diverse sensor modalities for activity recognition [Logan *et al.* 2007] many important issues have been identified that were not evident in prior work when data was collected under more controlled conditions.

Many applications in the field of medical diagnosis and elderly care require long-term activity monitoring for modeling and detecting changes in user behavior. Thus, long-term activity recognition [Huỳnh *et al.* 2007, Logan *et al.* 2007] has emerged as an important research topic in the activity recognition community aiming for recognition of daily routines [Huỳnh *et al.* 2008] or modeling of the users' rhythms [Van Laerhoven *et al.* 2008b]. Interestingly, an early attempt [Clarkson and Pentland 1999] to recognize low-level events such as walking into a building, crossing the street, or riding an elevator

and cluster them into high-level scenes such as shopping for groceries or going home dates back to 1999.

2.2 Health Care and Elderly Care Applications

Activity recognition enables a wide range of applications in the field of health care and elderly care.

In the health care domain a special focus is placed on encouraging physical activity [Consolvo *et al.* 2008a] among teenagers [Toscos *et al.* 2008], elderly [Albaina *et al.* 2009] or wheelchair users [Cuzzort and Starner 2008]. Many health care applications are disease dependent. For example, automatic assessment of spatio-temporal gait parameters [Salarian *et al.* 2004] is an important tool for estimation of motor function in Parkinson's disease patients. Similarly, automatic gait analysis [Lackovic *et al.* 2000] can be used to provide details on rehabilitation of orthopedic patients and their walking style patterns that might not be visible by visual observation. Recognition of self-stimulatory behaviors in autistic children [Westeyn *et al.* 2005] enables caregivers to explore the correlation between these behaviors and environmental factors or physiological markers. Detecting play activities [Westeyn *et al.* 2008] by augmenting toys with sensing capabilities allows for early identification of development delays in young children. Psychological studies of mood disorders benefit from automatic recognition of physical activities [Van Laerhoven *et al.* 2006] by correlating them, for example with mood swings in bipolar patients. There are also attempts to support sleep studies by detecting sleeping patterns [Van Laerhoven *et al.* 2008a] or sleep apnea [Oliver and Flores-Mangas 2007]. Recognition of wheelchair propulsion patterns [French *et al.* 2008] can help wheelchair users to learn patterns that are less damaging for upper limbs. Automatic dietary monitoring [Amft *et al.* 2007, Shroff *et al.* 2008] can improve the quality of life for diabetes patients and prevent obesity.

An important class of applications in the elderly care domain is detection of potentially dangerous situations. For example, if we were able to recognize that a person has fallen down [Degen *et al.* 2008, Jafari *et al.* 2007, Doukas and Maglogiannis 2008], an automatic emergency call system could be feasible. Considerable effort has been devoted to supporting people suffering from Alzheimer's disease through the reminders and memory aids [Backman *et al.* 2006, Lee and Dey 2008, Vurgun *et al.* 2007, Du *et al.* 2008]. Similarly, there have been efforts to support elderly people with cues on how to complete an activity [Mihailidis *et al.* 2003, Hoey *et al.* 2007, Wilson and Philipose 2005], or how to find a way to the final destination [Chang *et al.* 2008, Patterson *et al.* 2008, Liu *et al.* 2009a]. There are also attempts to deploy activity monitoring systems in elderly care facilities such as nursing homes [Allin *et al.* 2003, Hanser *et al.* 2008]. Applications that accumulate and summarize statistics about daily activities have been developed [Choudhury *et al.* 2006] for supporting ADL/IADL assessment.

2.3 Sensors and Multi-Sensor Approaches

In Chapter 1 we stated three important characteristics of daily activities, namely body-motion, usage of different objects, and location specificity for different activities. In order to capture these characteristics, different types of sensors have been proposed for activity recognition. Generally, they can be classified in two groups: 1) *wearable sensors* that are worn by users, and 2) *environmental sensors* that are deployed in the environment.

Assuming that motion of the body during the execution of an activity can robustly characterize the activity, wearable sensors strapped to the human body can be used to recognize the movement patterns while performing various activities. Typical sensors for these approaches are accelerometers [Bao and Intille 2004, Maurer *et al.* 2006, Ravi *et al.* 2005, Pärkkä *et al.* 2006]. They are unobtrusive, light-weight, and power-efficient. Moreover, research in the wearable computing community has shown that they lead to good recognition results. Another type of relatively simple and even more power efficient sensors are tilt switches [Van Laerhoven and Gellerson 2004]. They have also shown potential in capturing limited information about body motion. The combination of these two sensor modalities (i.e. accelerometers and tilt switches) has been successfully used in a more power efficient representation of human activity characteristics by switching between posture and motion capturing [Van Laerhoven *et al.* 2006]. An additional type of inertial sensors that is often used for capturing fine grained orientation is the gyroscope [Tanaka *et al.* 2004, Najafi *et al.* 2003]. Recently proposed magnetic sensors [Pirkl *et al.* 2008] also seem to be a promising approach to activity recognition.

Assuming that the used sequence of objects during the execution of an activity can robustly categorize the activity, different types of sensors deployed throughout the environment can enable detection of the objects people use. Examples of such sensors are RFID tags and readers [Philipose *et al.* 2004, Wyatt *et al.* 2005, Patterson *et al.* 2005, yau Lin and jen Hsu 2006], state-change switch sensors [Tapia *et al.* 2004], and wireless accelerometers that can be attached to the objects of interest to detect when they are being used [Tapia *et al.* 2007b].

Location has proven to be a good indicator for different activities, and numerous sensors have been used for this purpose, such as indoor motion detectors based on infra-red [Wren and Tapia 2006, Logan *et al.* 2007], ultra-sound [Ogris *et al.* 2005], and recently introduced infrastructure mediated [Patel *et al.* 2008] sensing or GPS sensors for outdoor location [Ashbrook and Starner 2003, Liao *et al.* 2005].

Furthermore, for activities with characteristic sounds, audio sensing [Chen *et al.* 2005, Maurer *et al.* 2006, Stäger *et al.* 2004, Choudhury and Pentland 2003] can be used to infer the user's activity. However, as microphones are often considered as being too intrusive, these approaches might have difficulties to become part of everyday practice. Similarly, the computer vision community is working in parallel on human activity recognition from video sequences (e.g. [Gavrila 1999, Mihailidis *et al.* 2004, Gao *et al.* 2004, Duong *et al.* 2005, Aghajan *et al.* 2007, Nowozin *et al.* 2007]). Also, wearable vision has gained

increasing attention (e.g. [Starner *et al.* 1998, Clarkson and Pentland 1999, Brashear *et al.* 2003]).

Heart-rate monitors [Chan *et al.* 2003, Tapia *et al.* 2007a], electrooculography [Bulling *et al.* 2008], and lately developed approaches for monitoring muscle activity [Lukowicz *et al.* 2006] and electrodermal activity [Westeyn *et al.* 2006, Schumm *et al.* 2008] or capacitive sensing [Cheng and Lukowicz 2008] are physiological sensors that can also potentially contribute to capturing certain aspects of activities such as their intensities, body stress, or mental activities.

In Chapter 1 it has been stated that capturing only one single characteristic of daily activities such as ADLs/IADLs might not be sufficient to enable robust detection in real-world settings. That has led to a few *multi-sensor approaches* aiming at exploiting different activity properties.

A dominant characteristic of many activities is body-motion. Therefore, there have been attempts to combine motion data with other sensor modalities, such as microphones (e.g. [Kern *et al.* 2004, Ward *et al.* 2005, Minnen *et al.* 2005, Lester *et al.* 2006]), wearable cameras [Brashear *et al.* 2003], location sensors [Stiefmeier *et al.* 2006, Subramanya *et al.* 2006] and recently, RFID tag readers [Wang *et al.* 2007, Stikic *et al.* 2008a]. Other types of multi-sensors approaches include for example combination of audio and video [Clarkson and Pentland 1999], RFID and video [Wu *et al.* 2007] or integrating multiple sensors on a single device [Van Laerhoven and Aronsen 2007, Choudhury *et al.* 2008].

In the past, researchers explored how the number of used sensors and their placement influences recognition performance. Using multiple accelerometers (up to 30 sensors in [Van Laerhoven and Gellerson 2004]) on different strategic body locations such as wrist, hip, or thigh (e.g. [Bao and Intille 2004, Lester *et al.* 2006, Huỳnh *et al.* 2007]) typically improves recognition results. However, wearing multiple sensors is often considered as too obtrusive and it would be highly desirable to decrease the number of the required sensors. In Chapter 4 we will show that a multi-sensor approach using complementary information from acceleration and RFID data is able to achieve high recognition scores with a small number of sensors. Furthermore, in Chapter 5 we explore the potential of multi-sensor approaches to decrease the level of supervision in activity recognition by fusing two complementary sensor modalities, i.e. wearable accelerometers and environmental infra-red motion detectors.

2.4 Annotation Techniques

Most activity recognition approaches rely on annotated activity data. We have argued in Chapter 1 that labeling of activities is one of the major challenges in the field of activity recognition. There exists a wide range of annotation techniques used for that purpose. An important difference between methods is whether they rely on offline annotations after the recording has been finished or they depend on online annotations during execution of

the activities. Both approaches have their advantages and drawbacks and in the following we will summarize them by giving a short review of typically used annotation methods for activity recognition studies.

A typical offline method is based on high-fidelity sensors, such as *audio and video recordings* of the activities for capturing fine-grained annotations. It has been applied e.g. in controlled and relatively short experiments [Stikic and Van Laerhoven 2007, Zinnen *et al.* 2007]. Recently, it has been used for a long-term activity study [Logan *et al.* 2007] in an instrumented home environment [Intille *et al.* 2006]. Even though this method provides accurate annotations, it is often feasible only in indoor environments and requires a significant amount of effort to annotate the data. In [Logan *et al.* 2007] an hour was spent on average for annotating 1.5 hours of data. Furthermore, this method is often not acceptable due to privacy concerns. Another offline methods include *subject's self-recall* [Van Laerhoven *et al.* 2008b] which might introduce significant noise due to recall-errors or *indirect observation* of sensor data [Tapia *et al.* 2004]. In the case of the environmental sensors such as RFID tags and switch sensors it might be relatively easy for a user to guess the activity from a sensor stream [Wilson *et al.* 2003], but in the case of wearable sensors such as accelerometers it might require an experienced user.

Online methods can be divided into two categories. The first category involves an external observer of the experiment and the second category requires a user himself to annotate the activities during the course of the study. The first group, so-called *direct observation* [Maurer *et al.* 2006], is suited only for controlled laboratory experiments as it scales poorly to a large number of users. Direct observation is even impossible in long-term studies where the goal is to capture real-world data of the user's typical daily activities under natural circumstances. Furthermore, restricted laboratory environments may artificially influence the way an activity is being performed. On the other hand, the *user-annotated* data can be provided in several ways, e.g. through *time diary* [Huỳnh *et al.* 2008] or *experience sampling* [Tapia *et al.* 2004, Froehlich *et al.* 2007]. When using the time diary annotation method, a user writes down the current activity together with the start and end time, either electronically or on paper. However, feasibility of this method highly depends on the type of activities, e.g. office related activities might be easily annotated this way, but for sport related activities this method would not be suitable. Furthermore, annotating data this way requires increased user awareness, and it might happen that the user forgets to annotate parts of the recorded data. One way of dealing with this issue is to provide a *script* of the activities a user needs to perform [Bao and Intille 2004] and after an activity is finished a user provides a timestamp and continues with the next activity in the script. This however may lead to staged activity recordings not reflecting natural ways activities are being performed in real-world settings. In order to overcome this issue the experience sampling method has been successfully used in other domains, in particular in psychology, for many years [Csikszentmihalyi and Larson 1987, Wheeler and Reis 1991, Hektner and Csikszentmihalyi 2002]. It aims to capture online annotations during recordings by periodical prompts of a user to provide information about his current activities. That way the user is reminded to annotate the data requiring less awareness and permitting better coverage. The method is fast and easy to use as it

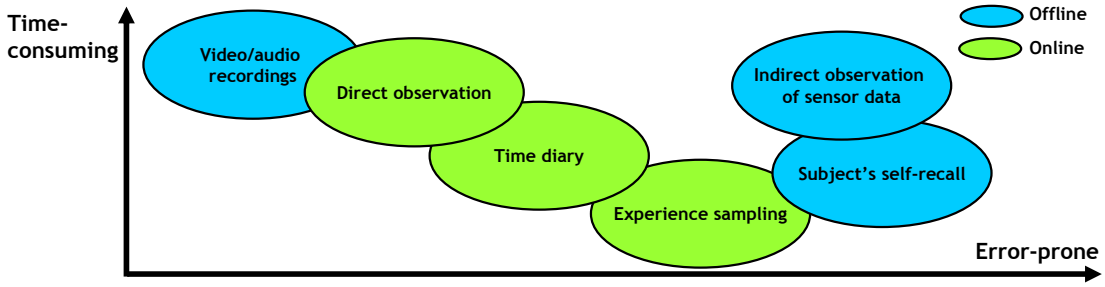


***Figure 2.1**: Sketch depicting the most commonly used annotation techniques for activity recognition, in function of how time-consuming and error-prone they tend to be.*

typically runs on a mobile phone of a user [Froehlich *et al.* 2007]. Furthermore, the prompts might be triggered at more appropriate times by special context-events [Intille *et al.* 2003]. Therefore, experience sampling has attracted a lot of interest not only in activity recognition but also generally in the ubiquitous computing community for evaluation of different applications and user studies [Consolvo and Walker 2003, Klasnja *et al.* 2008]. However, experience sampling can also be annoying for users especially when using a high sampling rate for capturing more detailed annotations.

Figure 2.1 illustrates a typical trade-off between accuracy of an annotation method and the time required for annotation. Methods that provide accurate annotations such as direct observations or video and audio recordings are labor-intensive, scale poorly to large numbers of users and activities, and are often not acceptable due to privacy concerns. In contrast, experience sampling and time diary require user involvement which can lead to recall errors and inaccurate annotations of short term activities, lack of temporal precision, and frequent interruptions that might change the activity itself and disrupt the user.

As experience sampling appears to be the most suitable method for long-term activity studies, in Chapter 6 and Chapter 7 we focus on decreasing the number of experience sampling prompts. For that purpose we introduce the algorithms that are capable of learning from very small amounts of labeled training data. Furthermore, we explore to which extent these methods can deal with ambiguously and incompletely labeled data, which is often the case in real-world recordings.

2.5 Algorithms for Activity Recognition

In this section we give a general overview of typical machine learning algorithms that have been applied for activity recognition. In Chapter 3 we will describe in more details the machine learning algorithms that are used in different parts of this thesis.

Most approaches for human activity recognition are based on state-of-the-art machine learning techniques. Most of the prior work relies on *supervised learning*, which requires *labeled* training data to train a classifier. Typically, the process consists of computing the

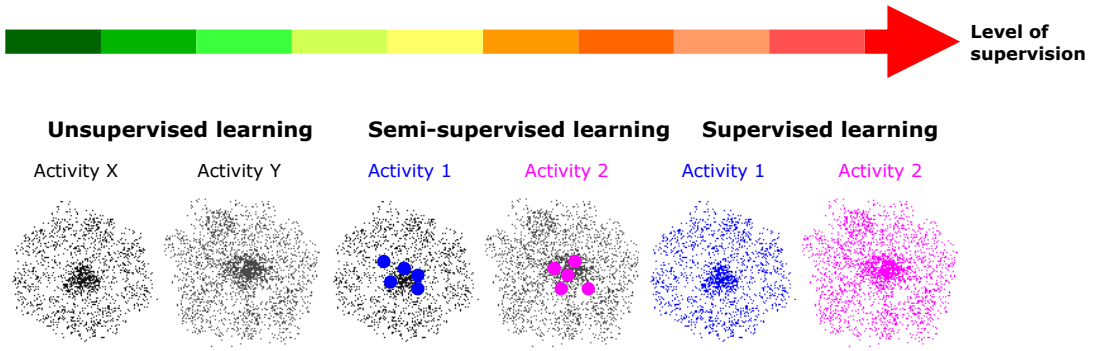


Figure 2.2: *Illustration of three different levels of supervision in activity recognition based on supervised learning, semi-supervised learning, and unsupervised learning. Supervised learning requires completely labeled training data and classifies activities, unsupervised learning does not need any labeled training data but it can not classify activities, and semi-supervised learning classifies activities based on a few provided labeled training data in addition to the large amount of unlabeled training data.*

set of *feature vectors* from the stream of sensor data over a sliding time window, feeding the extracted feature vectors into the algorithm to train a classifier, and then testing the trained classifier on an independent set of data.

The techniques can be categorized as either *generative* algorithms that model class-conditional distributions [Bao and Intille 2004, Ward *et al.* 2005] or *discriminative* algorithms that focus on learning the class decision boundaries [Ravi *et al.* 2005, Lester *et al.* 2006]. There also exist some hybrid approaches (e.g. [Lester *et al.* 2005, Huỳnh and Schiele 2006a]) that combine both generative and discriminative learning aiming to exploit the advantages of both techniques.

There is a wide range of the supervised classifiers that have been used for activity recognition such as Naive Bayes [Van Laerhoven *et al.* 2003, Tapia *et al.* 2004, Ravi *et al.* 2005], Decision Trees [Bao and Intille 2004, Maurer *et al.* 2006, Logan *et al.* 2007, Tapia *et al.* 2007a], Nearest Neighbor [Maurer *et al.* 2006, Van Laerhoven *et al.* 2008b], Hidden Markov Models [Lukowicz *et al.* 2004, Ward *et al.* 2005, Lester *et al.* 2006, Huỳnh *et al.* 2007], Support Vector Machines [Ravi *et al.* 2005, Huỳnh *et al.* 2007], and Boosting [Ravi *et al.* 2005, Lester *et al.* 2005, Minnen *et al.* 2007]. Recently, a string-matching method has also been proposed in [Stiefmeier *et al.* 2008]. Supervised algorithms typically achieve high recognition performance yet they require significant amounts of labeled activity data for training a classifier.

Moving beyond fully supervised settings, researchers have started studying the feasibility of other machine learning techniques to lower the annotation burden. Figure 2.2 illustrates two alternative approaches, namely *unsupervised learning* and *semi-supervised learning* and compare them to the standard supervised setting. The main difference between these three settings is whether they use labeled or unlabeled data, and if they are

capable of classifying activities in the end or not. Supervised learning uses only labeled training data to classify activities. Unsupervised methods do not require any labeled data. However, they cannot assign semantic meaning to the contexts they recognize. Semi-supervised methods aim at combining the advantages of both supervised and unsupervised learning by using only a small part of labeled training data in addition to a large amount of unlabeled data for training. In the end, the trained algorithm is able to classify activities. In the following we will describe these two techniques in more details.

Unsupervised learning. Unsupervised techniques enable discovery of structure in activity data without the need for labeled training data. In [Krause *et al.* 2003] the authors combined a Kohonen Self Organizing Map with k-means clustering in an online algorithm. In [Clarkson and Pentland 1999] a hierarchy of HMMs was used in an unsupervised way to cluster data. Previous work in motif discovery [Minnen *et al.* 2006a] and topic models [Huỳnh *et al.* 2008] focused on discovery and modeling of short-term motion primitives, and daily routines as a probabilistic combination of low-level activity patterns, respectively. In [Huỳnh and Schiele 2006b], an unsupervised algorithm based on multiple eigenspaces has been introduced. The approach is able to build low-dimensional models that correspond to different activities, without any prior training required. However, while the learned structure results in interesting representations of the data one still requires at least a few labels to achieve reliable classification results. There are also attempts to decrease labeling efforts by manually defining common sense models of daily activities [Wang *et al.* 2007] or mining these models from the web [Wyatt *et al.* 2005].

Semi-supervised learning. In a semi-supervised setting [Chapelle *et al.* 2006], typically there is only a small set of labeled training data available in addition to a substantial amount of unlabeled training data. In the context of activity recognition, there have been attempts to learn from both labeled and unlabeled data by using iterative algorithms such as self-training, co-training [Stikic *et al.* 2008b], and En-Co-Training [Guan *et al.* 2007]. These are wrapper algorithms that repeatedly use a supervised learning method to label part of the unlabeled training data. For graphical models it has been proposed to use virtual evidence [Subramanya *et al.* 2006] mechanism for depicting evidence in a Dynamic Bayesian Network (DBN) during joint reasoning about low-level activities and spatial context, as well as semi-supervised virtual evidence boosting [Mahdaviani and Choudhury 2007] for training Conditional Random Fields (CRF) on data collected from wearable sensors. Furthermore, in [Hoey *et al.* 2005] an expectation-maximization (EM) algorithm was used for semi-supervised learning of a partially observable Markov decision process (POMDP) for the task of handwashing.

Another alternative approach to reduce the labeling efforts is **active learning**. The goal here is to focus labeling efforts on the most profitable instances. Several active sampling functions have been proposed for activity recognition based on a multi-sensor approach [Stikic *et al.* 2008b] and for Hidden Markov Models [Anderson and Moore 2005]. Moreover, in an office-centered setting, different experience sampling strategies

based on active learning [Kapoor and Horvitz 2007] have been evaluated [Kapoor and Horvitz 2008] for predicting user's interruptibility.

The objective of this thesis is to make advances in semi-supervised learning and active learning for activity recognition. There exists relatively little work exploring semi-supervised techniques for human activity recognition. Furthermore, these approaches do neither address nor analyze the potential of multi-sensor approaches for the recognition of physical activities. Additionally, the evaluation of the proposed approaches was performed on relatively simplistic datasets consisting mostly of activities such as sitting, standing, walking, and running. In Chapter 5 we propose a multi-sensor semi-supervised approach based on co-training [Blum and Mitchell 1998] and evaluate it on a challenging publicly available dataset [Logan *et al.* 2007]. Furthermore, in contrast to previous work in semi-supervised learning for activity recognition, we propose and explore in Chapter 7 the use of graph-based semi-supervised techniques. These algorithms have proven to be powerful and versatile for different scenarios in machine learning. Further differences to previous work are that we use multiple graphs based on two different similarity measures for improved performance and that we also employ time as the basis for label propagation. On the other hand, the focus of active learning approaches is on the recognition of user's desktop activities for predicting interruptibility of a user. In Chapter 5 we make a first step towards active learning for physical activity recognition. Lastly, in Chapter 6 we explore a completely new direction for reducing the level of supervision in activity recognition based on multi-instance learning.

3

Methodology

As we have seen in Chapter 2, the field of activity recognition has progressed significantly in recent years. In order to further advance the state-of-the-art, we proceed in the following direction. For the evaluation of our algorithms we use realistic publicly available *datasets*, suitable for comparison of different approaches. The datasets range from a multi-person dataset of housekeeping activities to single-person but long-term daily activity datasets in non-laboratory settings. The goal of our *evaluation procedure* is to estimate whether the algorithms can generalize across multiple persons and over different days of activity recordings. As our algorithms employ different *classifiers*, we introduce the fundamentals of the classifiers used in the rest of the thesis.

The chapter is organized as follows. In Section 3.1 we describe the three datasets used for evaluation of our algorithms. Section 3.2 outlines several classifiers used in different stages of our algorithms. In Section 3.3 we present two different evaluation procedures and several figures of merit used in the following chapters.

3.1 Datasets

In this section, we describe the three datasets used in this thesis, namely:

- *Housekeeping*
- *PLCouple1*
- *TU Darmstadt*

The *Housekeeping* dataset is used in Chapter 4 for evaluation of the multi-sensor approach based on the combination of RFID and accelerometer sensing. The *PLCouple1* dataset is used in Chapter 5 for evaluation of the multi-sensor approach for reducing the level of supervision in activity recognition. Moreover, it is used in Chapter 6 and Chapter 7 for comparative evaluation of the multi-instance approach and the graph based semi-supervised approach, respectively. The *TU Darmstadt* dataset is also used in Chapter 6 and Chapter 7 for additional evaluation of these two approaches to avoid a potential bias to a single dataset.

Activity	Overall duration [min]	Average duration [min]
Vacuuming	21.7	3.6
Ironing	78.4	11.2
Dish washing	30.4	5.1
Dusting	21.4	2.4
Cleaning windows	36.6	5.2
Watering plants	4.2	0.5
Mopping	11.5	2.3
Brooming	11.9	2
Setting the table	10.5	1.8
Bed making	12.7	2.1

***Table 3.1**: Housekeeping dataset: Overall and average duration of activities.*

***Figure 3.1**: Subject performing different housekeeping activities.*

3.1.1 Housekeeping Dataset

This dataset has been recorded at TU Darmstadt. The focus of the experiment is on one specific class of IADLs, i.e. *housekeeping* activities. In our activity recordings, we target the following housekeeping activities: *vacuuming, ironing, dish washing, dusting, cleaning windows, watering plants, mopping, brooming, setting the table, and bed making*. The overall length of the dataset is 240 minutes. As can be seen in Table 3.1, the overall and average duration strongly varies among the activities. The dataset is publicly available [Stikic *et al.* 2008a].

The goal of the experiment is to explore the combination of two important activity characteristics (see Section 1.1): *body motion* and usage of different *objects* during the execution of activities. As housekeeping activities (Figure 3.1) typically require distinctive hand movements, we limit the hardware to wrist-worn sensors. That way we are able to capture the key person-object interactions and movements during activities and still have a satisfactory level of the wearability and unobtrusiveness. Figure 3.2 shows the

***Figure 3.2**: Wearable sensors used in the experiment.*

sensors used for recording the dataset. The subjects wore the sensors on their dominant wrist. We use a single 3D accelerometer to infer relevant arm movements. The person-object interactions are detected with a wrist-worn RFID reader. The whole experiment is recorded by video camera for offline annotations of activities, to avoid a potential bias in the dataset by the subjects' online annotations.

We had 12 subjects participating in the experiment, 3 females and 9 males, including 3 left-handed subjects. As we wanted to avoid biasing the dataset, the scenario presented to the subjects was kept as vague as possible. The subjects were told to choose a certain set of activities to perform based on the list of 10 targeted activities. We specifically did not give a detailed description of the required object interactions and sequences of actions to be performed within each activity, as we wanted to avoid biasing our dataset with same sequences of tagged objects and artificially staged actions. The subjects were encouraged to perform the activities as natural as possible and using their own routines as much as possible, resulting in a wide variety of ways different people performed the same activity. We will address this issue and the identified challenges in a series of our data recording experiments in Chapter 4.

3.1.2 PLCouple1 and TU Darmstadt Datasets

The PLCouple1 and TU Darmstadt datasets are publicly available datasets recorded over longer periods of time in real-world settings comprising typical non-scripted daily activities of a single subject. What makes these datasets especially challenging is the fact that the amount of data for activities varies a lot for different activities reflecting the natural distribution and duration of activities in real life. The datasets include fine-grained annotations of activities which make them suitable for systematic analysis of different activity recognition approaches. Here, we outline the main characteristics of the datasets that are of interest for the rest of the thesis.

Activity	Overall duration [min]	Average duration [min]	Min. duration per day [min]	Max. duration per day [min]
Watching TV or movies	732.3	33	0	180.3
Dish washing	9.6	1.25	0	5.2
Eating	222.9	2.5	0	55.5
Grooming	39.7	2.25	0.2	18.7
Hygiene	39.4	2.75	0	9.5
Meal preparation	41.6	1.75	0	21.4
Using computer	1500.4	18.75	4.6	288.1
Using phone	154.5	3.25	0	57.5

***Table 3.2**: PLCouple1 dataset: Overall and average duration of activities and their minimum and maximum daily duration.*

PLCouple1. The PLCouple1 [Logan *et al.* 2007] dataset is recorded at the PlaceLab [Intille *et al.* 2006], a highly instrumented home environment, where a couple moved in and lived there for 10 weeks, continuing as normal a routine as possible. An audio-visual recording system was used for capturing ground truth and an expert annotated 104 hours of the male's activities, comprising data collected on 15 separate days. In our experiments, we use a publicly available subset of 68 hours of annotated data collected on 9 separate days. Despite a substantial amount of data collected and annotated, there is still a lack of data for many fine-grained activities, which led to 9 activities to be studied in [Logan *et al.* 2007]. In this thesis we focus on the same set of activities: *actively watching TV or movies*, *dish washing*, *eating*, *grooming*, *hygiene*, *meal preparation*, *reading paper/book/magazine*, *using computer*, and *using phone*. All other activities in the dataset are considered as an *unknown* class. Table 3.2 shows different statistics of the activities such as overall and average duration as well as their minimum and maximum daily duration.

The PlaceLab facility contains over 900 sensors and the goal of the experiment in [Logan *et al.* 2007] was to compare different sensor modalities under the same real-world conditions. Motion sensors, namely body-worn accelerometers [Tapia *et al.* 2006] and infra-red sensors [Wren and Tapia 2006] outperformed other sensors (i.e. RFID and environmental built-in sensors). The male subject wore 3 3D accelerometers on the dominant wrist, the dominant hip, and the non-dominant thigh. For the recordings, the sampling frequency of $20Hz$ was used. Ten wireless infra-red sensors were installed around the apartment to detect motion in each room: bedroom, bathroom, powder room, office, office hallway, kitchen, kitchen hallway, foyer, living room, and dining room. As the dataset was recorded only at times when the subject was at home, there is a certain number of gaps in the data. Additionally, there are also a few gaps due to wireless communication.

In Chapter 5 we use data from both sensor modalities (i.e. accelerometers and infra-red sensors) in order to reduce the required level of supervision. In Chapter 6 and Chapter 7 only acceleration data is used, enabling activity recognition without any external infrastructure needed.

Activity	Overall duration [min]	Average duration [min]	Min. duration per day [min]	Max. duration per day [min]
Sitting/desk activities	3077.8	434	331.9	496.8
Lying/reading/using computer	201.4	99.7	0	152.5
Having dinner	127.8	18.5	7.2	28.1
Walking freely	126.7	18.5	10.7	23.8
Driving car	122.8	24.7	0	34.6
Having lunch	76.6	11.2	4.8	13.2
Discussing at whiteboard	63.9	32	0	36.3
Driving bike	47.3	23.7	0	24.8
Standing/talking on phone	25.2	6.7	0	20.3
Walking/carrying something	23.5	4.3	0	7.6
Walking	23.3	4.2	0	8.9
Picking up cafeteria food	23.1	3.7	2	5.3
Sitting/having a coffee	22.2	6	0	7.5
Queuing in line	20.2	4.5	0	8.7
Personal hygiene	17.5	4.7	0	6.6
Using the toilet	17.1	3.7	0	6.3
Washing dishes	13.1	3.7	0	5.3
Brushing teeth	4.4	1.5	0	2
Standing/using the toilet	3.1	1	0	1
Washing hands	2.2	1	0	1.1

***Table 3.3**: TU Darmstadt dataset: Overall and average duration of activities and their minimum and maximum daily duration.*

TU Darmstadt. The TU Darmstadt dataset [Huỳnh *et al.* 2008] consists of data from 2 3D accelerometers worn on the dominant (right) wrist and in the right hip pocket recorded during a period of 16 days. Data from 7 days are annotated by a test subject combining different online and offline annotation methods, resulting in 84 hours of usable annotated data. Due to the memory constrains (512kb) of the sensor platform, data had to be recorded at a relatively low frequency of 2.5Hz. Still, the memory had to be emptied every 4 hours, producing a certain number of gaps in the data.

The dataset contains 34 distinct activities. Since many activities appeared only on one day during the recorded time period, we could use only a subset of 20 different activities in our experiments due to our evaluation procedure (see Section 3.3). Therefore, we target the following activities along with the *unlabeled* class: *sitting/desk activities*, *lying while reading/using computer*, *having dinner*, *walking freely*, *driving car*, *having lunch*, *discussing at whiteboard*, *driving bike*, *standing/talking on phone*, *walking while carrying something*, *walking*, *picking up cafeteria food*, *sitting/having a coffee*, *queuing in line*, *personal hygiene*, *using the toilet*, *washing dishes*, *brushing teeth*, *standing/using the toilet*, and *washing hands*. Table 3.3 shows overall and average activity durations as well as their minimum and maximum daily duration.

3.2 Classifiers

In this section we shortly outline the fundamentals of the classifiers used in this thesis:

- *Naive Bayes*
- *Hidden Markov Models*
- *Joint Boosting*
- *Decision Trees*
- *Support Vector Machines*

In Chapter 4 we use Naive Bayes, Hidden Markov Models, and Joint Boosting for classification of acceleration activity data. In Chapter 5 we use the Naive Bayes, Decision Trees, and Joint Boosting classifiers for the supervised analysis of activity data and we also employ Joint Boosting as the underlying classifier in our semi-supervised and active learning algorithms. In Chapter 6 and Chapter 7 we use the SVM classifier as a supervised baseline.

3.2.1 Naive Bayes

Naive Bayes is a supervised learning algorithm that requires labeled training data. It is a simple yet effective generative classifier based on *Bayes' theorem*:

$$p(y_j|x) = \frac{p(x|y_j)p(y_j)}{p(x)} \tag{3.1}$$

where $p(y_j|x)$ is the posterior probability of a class $y_j, j \in \{1,...,C\}$ given an n-dimensional feature vector $x = (x_1,...,x_n)$. $p(y_j)$ is the class prior probability, provided either a priori or estimated from training data. $p(x|y_j)$ is the *likelihood* that can be calculated from training data as:

$$p(x|y_j) = \prod_{i=1}^{n} p(x_i|y_j) \tag{3.2}$$

assuming that the different components x_i of the feature vector x are independent of each other. Even though the Naive Bayes classifier assumes that the components of a feature vector are independent, it often outperforms more sophisticated classifiers. $p(x)$ is used for normalization of relative likelihoods in order to represent absolute probabilities (i.e. $\sum_{j=1}^{C} p(y_j|x) = 1$):

$$p(x) = \sum_{j=1}^{C} p(x|y_j)p(y_j) \tag{3.3}$$

For classification, each feature vector x is assigned the label $\hat{y}$ of the class that has the highest posterior probability $p(y_j|x)$ (Equation 3.1):

$$\hat{y} = \underset{j\in\{1,..,C\}}{\operatorname{argmax}} \left(p(y_j|x)\right) \tag{3.4}$$

For probability density estimation of $p(x|y_j)$ different methods can be used. In Chapter 4 and Chapter 5, we employ the *unimodal Gaussian* model for continuous data (i.e. acceleration). In case of discrete sensor events such as infra-red motion detection firings we apply two different generative models [McCallum and Nigam 1998]: *multinomial* model when using the number of the activations as a feature and the *multi-variate Bernoulli* model for binary features.

3.2.2 Hidden Markov Models

Hidden Markov Models (HMM) belong to the class of statistical models aiming to capture the temporal structure of a signal. An important characteristic of the model is that the *state* is not directly observable, but the visible *observation* is a probabilistic function of the state. The model is defined by its set of parameters:

$$\lambda = (A, B, \pi) \tag{3.5}$$

where A is the set of state transition probabilities, B is the set of observation probabilities in a certain state, and π is the initial state distribution. Furthermore, by defining the number of states N and the number of distinct observation symbols M the model is completely specified.

During HMM training the goal is to adjust the model parameters $\lambda = (A, B, \pi)$ to best describe how the training observation sequence $O = O_1 O_2 ... O_T$ is generated, i.e. to maximize $P(O|\lambda)$. For that purpose typically the *Baum-Welch* algorithm is applied, which is an iterative *Expectation-Maximization (EM)* procedure for finding a local maximum of $P(O|\lambda)$. Further details about the algorithm can be found in [Rabiner 1989]. In Chapter 4 we use HMM with continuous acceleration signal. For that purpose, the probabilities B of discrete symbols have to be replaced by a probability density function, typically by a mixture of Gaussians. Once we have the trained models for each class, i.e. activity of interest, the classification of test observation sequence is performed by computing the likelihood of each model $P(O|\lambda)$ based on the *forward-backward* procedure [Rabiner 1989] and assigning it the label of the model with the highest likelihood. In the experiments, we use the HMM Toolbox for Matlab [Murphy 1998].

3.2.3 Joint Boosting

Joint Boosting [Torralba *et al.* 2004] is a multi-class variant of traditional boosting approaches. In standard boosting multiple weak learners $h_m(x)$ are combined into a single strong classifier H:

$$H(x) = \sum_{m=1}^{M} h_m(x) \tag{3.6}$$

Each weak learner $h_m(x)$ is a decision or regression stump on a single component x_i of a feature vector $x = (x_1, ..., x_n)$:

$$h_m(x) = a\delta(x_i > \theta) + b \tag{3.7}$$

where θ is the optimal threshold being automatically found, and δ is the indicator function being 1 or 0, depending on the condition $x_i > \theta$. Regression parameters a and b intuitively represent the confidence in judging a sample as the positive or negative class, respectively.

Joint Boosting is especially appealing because it finds the features that can be shared across the classes, which results in a faster classifier that needs less features than standard approaches. At each boosting round different subsets of classes $S \subseteq \{y_1, ..., y_C\}$ are examined for fitting a weak learner to distinguish that subset of classes from the other classes. The subset S_m that maximally reduces the error on the *weighted* training set for all the classes is chosen. The best weak learner is then shared among the classes in that subset:

$$H_{y_c}(x) = \sum_{m=1}^{M} h_m(x)\delta(y_c \in S_m) \tag{3.8}$$

Each training sample is assigned a weight w_i which enables focusing the training procedure on harder samples. In each boosting round these weights are updated by increasing the weights of the samples that are misclassified and decreasing the weights of the samples that are correctly classified:

$$w_i = w_i e^{-z_i^{y_c} h_m(x)} \tag{3.9}$$

where $z_i^{y_c} \in \{-1, +1\}$ are the membership labels for class y_c. For classification, each feature vector x is assigned the label $\hat{y}$ of the class that has the highest confidence prediction score (Equation 3.8), i.e.:

$$\hat{y} = \underset{j \in \{1,..,C\}}{\operatorname{argmax}} \left(H_{y_c}(x)\right) \tag{3.10}$$

Further details about the regression parameter estimation and the search heuristic for finding the best sharing subset of classes can be found in [Torralba *et al.* 2004]. Namely, the exhaustive search over all possible subsets of classes is not feasible due to the complexity $O(2^C)$. Thus, a greedy best first search strategy in applied to obtain an approximation of the best sharing subset of classes, reducing the complexity to $O(C^2)$.

3.2.4 Decision Trees

Decision tree learning is based on inductive inference in which the learned classification function is depicted by a decision tree. Generally, decision trees represent a disjunction of conjunctions of constraints on the components of a feature vector, enabling their representation as a set of *if-then* rules. Each node in the tree specifies a test of a feature vector component and each branch descending from that node corresponds to one of the possible test outcomes. The leaf nodes are associated with the set of all possible classes, and test

data are classified by sorting them down the tree from the root to the appropriate leaf node. The training phase consists of building a tree top-down, i.e. choosing which component of the feature vector should be tested at each node of the tree based on the *information gain* [Quinlan 1993], i.e. the expected reduction in entropy caused by partitioning the training data according to that feature.

In the experiments in Chapter 5 we employ the *C4.5* variant of a decision tree algorithm found in the Weka Machine Learning Algorithms Toolkit [Witten and Frank 2005]. It supports continuous data such as acceleration. Furthermore, it can successfully cope with overfitting by the *post-pruning* step that removes the nodes from the tree if the estimated accuracy is increased that way.

3.2.5 Support Vector Machines

Support Vector Machines (SVM) [Vapnik 1998] belong to the standard supervised linear binary classifiers. Here, we briefly outline the fundamentals of SVM classification. Further details can be found in e.g. [Burges 1998]. The goal is to construct an optimal separating hyperplane:

$$\mathbf{w}\cdot\mathbf{x}+b=0 \tag{3.11}$$

in the feature space of labeled training data $(\mathbf{x_i}, y_i)$, $\mathbf{x_i}\in\mathbb{R}^{\mathbf{N}}$, $y_i\in\{-1,+1\}$, $i=1,...,l$ that minimizes the expected generalization error. This is done by maximizing the *margin*, i.e. the distance from the hyperplane to the nearest data points (so-called *support vectors*):

$$\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2,\quad s.t.\quad y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1,\ \ i=1,\ldots,l \tag{3.12}$$

In the case of *non-separable* classes, soft-margin is maximized by introducing *slack variables* ξ_i that allow misclassification of training data:

$$\min_{\mathbf{w},b,\xi}\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{l}\xi_i,\quad s.t.\quad y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1-\xi_i,\ \ \xi_i\geq 0,\ \ i=1,\ldots,l \tag{3.13}$$

where C is a misclassification penalty parameter that controls the trade-off between training error and margin (larger C corresponding to a higher penalty for misclassifications). A *nonlinear* classifier is obtained by using a *kernel* transformation, i.e. the data are implicitly mapped to a high dimensional feature space where it is more likely that the two classes are linearly separable. In our experiments we employ the Gaussian radial basis function (RBF) kernel:

$$K(\mathbf{x_i},\mathbf{x_j})=e^{-\gamma\|\mathbf{x_i}-\mathbf{x_j}\|^2} \tag{3.14}$$

and SVM^{light} [Joachims 1999] implementation of an SVM learner that enables training on large sets of data. In order to extend the binary SVM classifier to our multi-class activity recognition setting in Chapter 6 and Chapter 7, we apply the typical "one vs. rest" approach by training N SVMs, each separating a single class from all remaining classes. During classification, each test sample is assigned the class of the SVM classifier which provides the highest prediction score for that sample.

3.3 Evaluation Procedure

This section defines the evaluation procedures and different figures of merit used in the experiments in the following chapters.

3.3.1 Cross-validation

Cross-validation is a technique for estimating the generalization ability of a classifier on an independent dataset. Typically, only a limited amount of data is available and cross-validation enables to predict the classifier's performance in practice by partitioning a dataset into *training set* and *test set* and performing multiple rounds by using different partitions. The validation results are then averaged over the cross-validation rounds.

Leave-one-person-out. In case of the multi-person *Housekeeping* dataset, we aim to examine the feasibility of person-independent activity recognition. For that purpose, we perform *leave-one-person-out* cross-validation on the data in the following manner. In each cross-validation round, we use data from all but one person for training. The classifier is then tested on the left out persons' data. The procedure is iteratively repeated for all subjects in the experiment.

Leave-one-day-out. In case of long-term activity recordings over multiple days (i.e. the *PLCouple1* and *TU Darmstadt* datasets) we conduct the experiments in a *leave-one-day-out* cross-validation manner to generate independent test data. In each cross-validation round we use one day of data for testing and the data from other days for training. The procedure is then repeated until data from all days have been tested.

		Classification	
		Positive	**Negative**
Ground truth	**Positive**	**TP**	**FN**
	Negative	**FP**	**TN**

***Figure 3.3**: Confusion matrix*

3.3.2 Evaluation Criteria

As figure of merit we use the following measures: *precision*, *recall*, *accuracy*, and *the area under the Receiver Operating Characteristic (ROC) curve (AUC)*. In our experiments

we deal with the *multi-class* problem, i.e. we aim to recognize different activity classes ($n > 2$). For simplicity, Figure 3.3 shows 4 possible outcomes of the *binary* classification task in a form of a *confusion matrix*:

- true positives (TP) – correctly classified positive examples
- true negatives (TN) – correctly classified negative examples
- false positives (FP) – misclassified negative examples
- false negatives (FN) – misclassified positive examples

that are used for defining the used figures of merit.

Precision is the percentage of positive predictions that are correct, i.e. the number of true positives in the test set divided by the sum of true positives and false positives in the test set:

$$Precision = \frac{TP}{TP+FP} \tag{3.15}$$

Recall is the percentage of positive examples that are correctly classified, i.e the number of true positives in the test set divided by the sum of true positives and false negatives in the test set:

$$Recall = \frac{TP}{TP+FN} \tag{3.16}$$

Accuracy is the number of correctly classified samples divided by the number of all test samples:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.17}$$

The area under the ROC curve provides an overall goodness measure of a classifier. The ROC curve plots true positive rate (TPR) vs. false positive rate (FPR) at all possible classifier's thresholds:

$$TPR = \frac{TP}{TP+FN} \tag{3.18}$$

$$FPR = \frac{FP}{FP+TN} \tag{3.19}$$

In the rest of the thesis, we mostly use accuracy extracted from multi-class confusion matrices, since it is a more intuitive and often used measure for the multi-class activity recognition settings. Exceptionally, in Chapter 4 precision and recall are used for measuring the performance of an RFID classifier. In Chapter 5 area under the ROC curve is used for comparison of the results reported in [Logan *et al.* 2007].

4

Combination of RFID and Accelerometer Sensing

RFID tag readers and accelerometers are two sensing technologies that have recently dropped in both size and cost. Assuming that key household items can easily be tagged, one could legitimately imagine a wrist-worn device which incorporates both to infer Activities of Daily Living. This chapter presents an effective and unobtrusive activity recognition system based on the combination of these two sensor modalities. We evaluate our algorithms on non-scripted datasets of 10 housekeeping activities performed by 12 subjects. We analyze different acceleration features and algorithms, and by analyzing tag detections we suggest the best tags' placements and the key objects to be tagged for each activity. The experimental results show that sensor fusion allows to compensate for the shortcomings of both sensor modalities while significantly improving recognition accuracy.

4.1 Introduction

In Chapter 1 we stated three important characteristics of daily activities: body motion, interactions with objects, and location. This chapter explores the first two characteristics in more details. As we have seen in Chapter 2, research in the wearable computing community has shown that characteristic movement patterns for different activities can be inferred from body-worn accelerometers. Assuming that the objects people use during the execution of an activity can also robustly categorize the activity, one can place sensors in the environment to detect user's interactions with objects. RFID tags and readers are typically used for that purpose due to their durability, small size, and low costs. We start from the hypothesis that the recognition results can be significantly improved by using both sensor modalities.

The goal of the research presented in this chapter is to improve the recognition results by integrating these two approaches, while also aiming to compensate for the shortcomings of both. In order to be able to accurately recognize different activities, the RFID

approach requires a large number of objects to be tagged. However, we argue that it is not feasible to tag all objects, because of several reasons. First, the deployment of large numbers of tags is still time consuming and error prone (see Section 4.2.2). Second, it is not practical to tag some objects because of their material (e.g. metal) or specific usage (e.g. objects used in microwave). We propose to use only the *key objects* for a specific set of activities by augmenting the object usage with a complementary sensing technique (i.e. accelerometers). On the other hand, accelerometer approaches often use multiple sensors placed on strategic body locations, such as wrist, hip, and thigh for accurate recognition. We propose to use only a single 3D accelerometer at the dominant wrist of the user, since limiting the hardware to a single wrist-mounted device containing both the RFID tag reader and the accelerometer could increase user acceptance of the automatic activity monitoring system.

The first contribution of this chapter is the combination of RFID and accelerometer sensing into an integrated activity recognition scheme that yields better recognition scores than either sensing technology alone. The second contribution are experimental results with different number of tagged objects which show that satisfactory recognition results can be achieved with fewer tagged objects than are typically used. The third contribution is a detailed analysis of different ways of combining the activity recognition results from the two sensor modalities as well as an evaluation of different features and window lengths. Moreover, the algorithms are evaluated on a challenging multi-person *Housekeeping* dataset introduced in Section 3.1.1.

The chapter is organized as follows. In Section 4.2 we further explain the critical choices made in a series of our *Housekeeping* dataset recordings. Section 4.3 presents an initial analysis of the recorded activity data. Section 4.4 introduces the three activity recognition approaches motivated by the initial data analysis. In Section 4.5 we report on the results of all three used approaches. Finally, in Section 4.6 we briefly summarize our results.

4.2 Experiment Setup

For our recordings we used a controlled lab environment (Figure 4.1) that was converted in a living space by furnishing the laboratory with typical objects found and used in a domestic setting, to make it resemble a common home environment. None of the subjects felt uncomfortable performing the home activities at the laboratory or wearing the sensors during the execution of the activities.

4.2.1 Hardware Setup

We use two types of wearable sensors, called *iBracelet* and *Porcupine*, for detecting object usage and arm movements, respectively. Both these sensors have been designed to be as

***Figure 4.1**: Laboratory where the dataset is recorded.*

unobtrusive and easy to deploy as possible. They also operate without any calibration requirements, making them an ideal solution for long-term monitoring necessary for the detection of changes in human behavior.

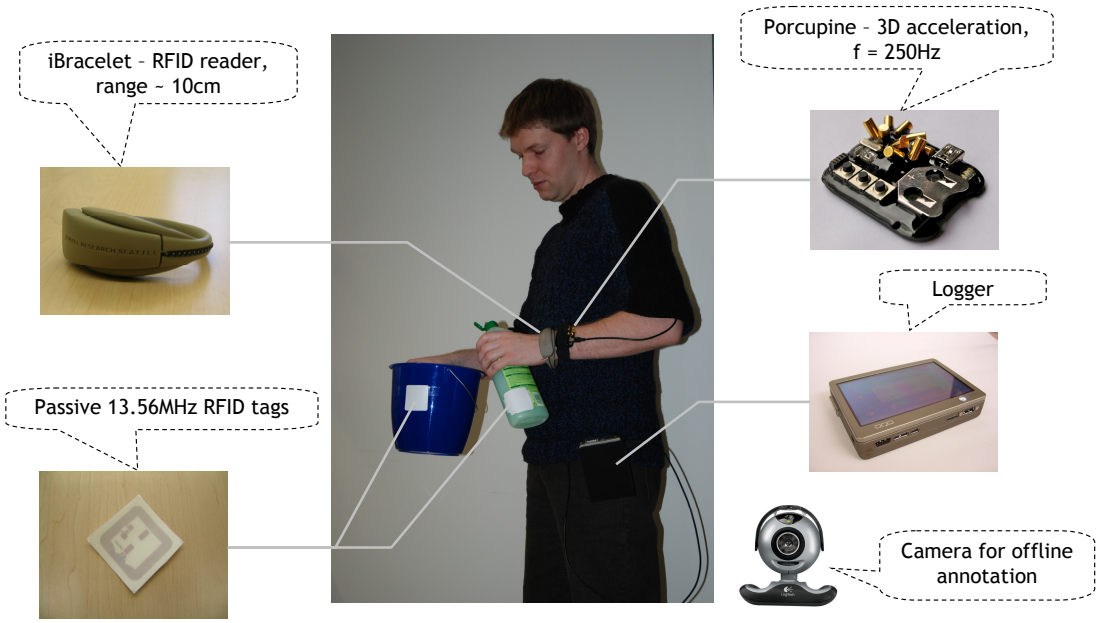


***Figure 4.2**: Hardware setup*

The iBracelet is a wrist-worn RFID reader (Figure 4.2) built at Intel Research Seattle [Fishkin *et al.* 2005] that can detect 13.56MHz RFID tags in the range of up to 10cm. We use passive adhesive 55 x 55 mm RFID tags (Figure 4.2). When queried by a reader, tags respond with a unique identifier. The bracelet wirelessly transmits the tag ID to a base station. The received "tag read" events are stored together with a timestamp to the persistent memory for later analysis. Therefore, if we tag the objects of interest, we can easily infer person-object interactions.

The RFID tags are durable, small sized and inexpensive. The RFID technology has gained growing interest in recent years, due to its potential in supply chain management.

Thus, in the near future, consumer goods could be permanently embedded with RFID tags, which would make the deployment of this technology feasible in home environments. Meanwhile, controlled deployment might still be feasible in the environments of the elderly, where they live semi-independently with occasional help of care givers who could add tags to accommodate the RFID monitoring.

The Porcupine is a wearable multi-sensor platform (Figure 4.2) developed at TU Darmstadt [Van Laerhoven and Aronsen 2007], which includes the following sensors: 3-axis accelerometer, 9 tilt switches, 1 temperature sensor, and 2 ambient light sensors. Additionally, it has a real-time clock, serial flash memory, mini USB port, 3 buttons for annotations, 3 LEDs and automatic power switching between battery and USB powered modes.

The Porcupine is attached to the wrist of the users by a strap in order to infer the fine grained arm movements. In principle, it could be easily integrated in a bracelet or watch to make it more unobtrusive for a user. For our recordings, we aim to have the highest and most accurate sampling frequency ($f = 250Hz$) for raw acceleration data. So we use the Porcupine in the USB powered mode by attaching it to a hip-worn OQO pocket-sized computer (Figure 4.2) that we also use for logging the data.

***Figure 4.3**: Tagged objects*

Tagging the objects. We deployed 191 tags on 58 objects (Figure 4.3). The number of tags per object varies between 18 tags for a pillow and 1 tag for a dusting cloth (Table 4.1). On average, 3 tags were deployed per object. Optimizing for detection, we aimed to tag various parts of the objects that we considered being hard for the interaction detection due to their size, the way they are usually being used by users, and the short range of the RFID reader antenna embedded in the iBracelet.

Multiple tags have been deployed on as many objects as possible for several reasons: 1) to find the key objects for the targeted set of housekeeping activities, 2) to evaluate the influence of the number of deployed tags on the recognition results by using different number of tags, and 3) to optimize tag detection for objects that are difficult to detect because of their size, shape or material.

Tagged objects	Number of tags per object
Pillow	18
Vacuum cleaner, bucket, bed sheet, blanket, table sheet	9
Mop, TV, big broom	7
Windows 1-4	7 each
Small pillow case, big pillow case	6
Ironing board	5
Sewing machine, dusters box	4
Water tap	3
Fan, iron, window cleaning liquid, pot for distilled water, bottle for distilled water, cleaning cloth, dish washing liquid, cupboard	2
Plates 1-4	2 each
Sewing machine utensils box, sewing machine mechanism box, dust pan, sponge, glove for left hand, glove for right hand, small broom, squeegee, watering can, water spray, flower pot	1
Dusting cloths 1-8 Wall sockets 1-4 Glasses 1-4	1 each

***Table 4.1**: Number of tags per object.*

4.2.2 Deployment Issues

In terms of the deployment, we faced some difficulties during the process of tagging a large number of objects and manually mapping the tag ID to the object it has been attached to. This process is tedious and a few errors occurred during the cataloging procedure. The errors were discovered afterwards during the analysis of the recorded data. This might not be an issue if in the future objects we buy are already equipped with RFID tags.

Another problem that appeared during the recordings was synchronization of data coming from the iBracelet, the Porcupine, and the video camera, which all produce data at different speeds. Data streams coming from the iBracelet and the Porcupine are timestamped, so their synchronization was almost straight forward, but yet, a significant time drift occurred now and then. We successfully recovered the data by identifying the time drift from the recorded videos. For that reason, in the later stage of the experiment we switched to logging both data streams on the same computer and performing the sensor fusion by logging the data in the same log file. Additionally, a test tag was used at the beginning of the recording for synchronization. We also asked the subjects to make 3 repetitive arm movements before starting with an activity for the synchronization of the video stream with the rest of the data. These movements are represented in the raw acceleration data by 3 distinguishable peaks, and can be easily synchronized by visual inspection of the acceleration data.

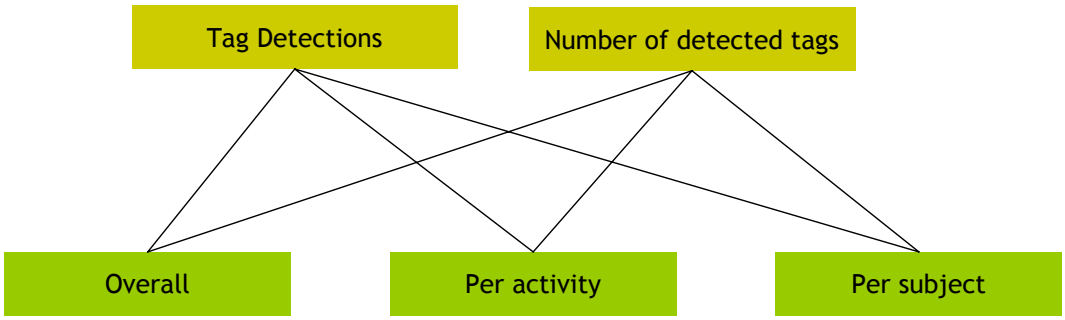


***Figure 4.4**: Three different kinds of analysis of the RFID data.*

4.3 Initial Analysis

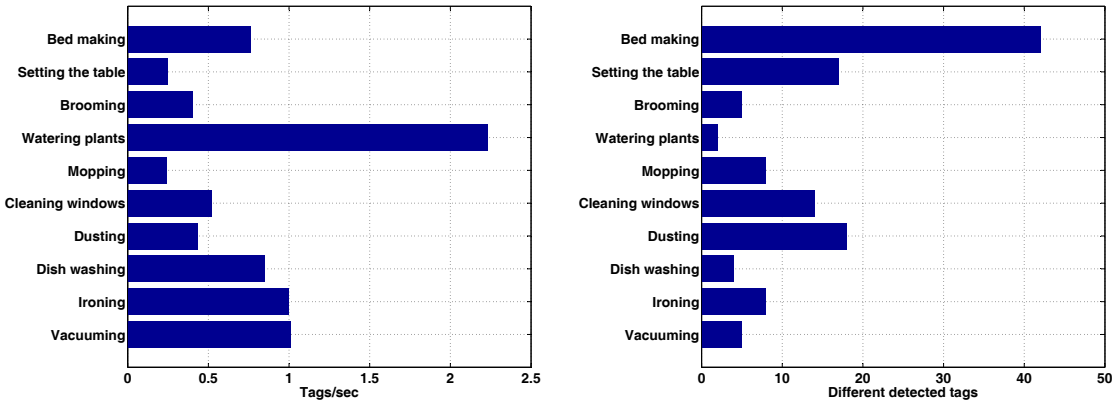
In order to provide a better insight into the advantages and shortcomings of the RFID and acceleration data, we will first present an initial analysis of the *Housekeeping* dataset introduced in Chapter 3.

4.3.1 Reliability of Tag Detection

Regarding the reliability of the RFID tag detection, we analyze the logged tag IDs in several ways (Figure 4.4). We compute the number of tag detections and different detected tags in three different ways: 1) *overall* in the dataset, 2) for each *activity*, and 3) for each *subject* in the experiment.

Throughout the whole dataset, the tags were detected on 10998 occasions. On average, 0.76 tag detections happened per second. Overall, the RFID reader detected 114 different tags, which is not much considering the 191 deployed tags. Two reasons can be found from the video footage. On the one hand, the short range of the RFID reader caused many false negatives, i.e. tags were not detected even though the subjects were interacting with the tagged objects. On the other hand, some of the tagged objects were not used during the execution of the recorded set of activities. Interestingly, in the whole dataset only 16 false positive readings (i.e. the number of tags detected accidentally near the hand) occurred.

Distribution of tag detections over the recorded activities is shown in Figure 4.5(a). The watering plants activity is the most extensive in terms of the tag detections (2.23tags/sec). This is due to the fact that the tags placed on the water spray and the watering can are often detected during watering plants. A similar situation occurs with the tags placed on the vacuum cleaner stick and the handle of the iron. That is the reason why the vacuuming and ironing activities also have very high tag detection scores (1.01tags/sec and 0.99tags/sec, respectively). However, as the tags placed on the mop, the broom and the dusting cloths are rarely detected, the mopping, brooming and dusting activities have lower tag detection scores (0.24tags/sec, 0.4tags/sec, and 0.43tags/sec, respectively). We



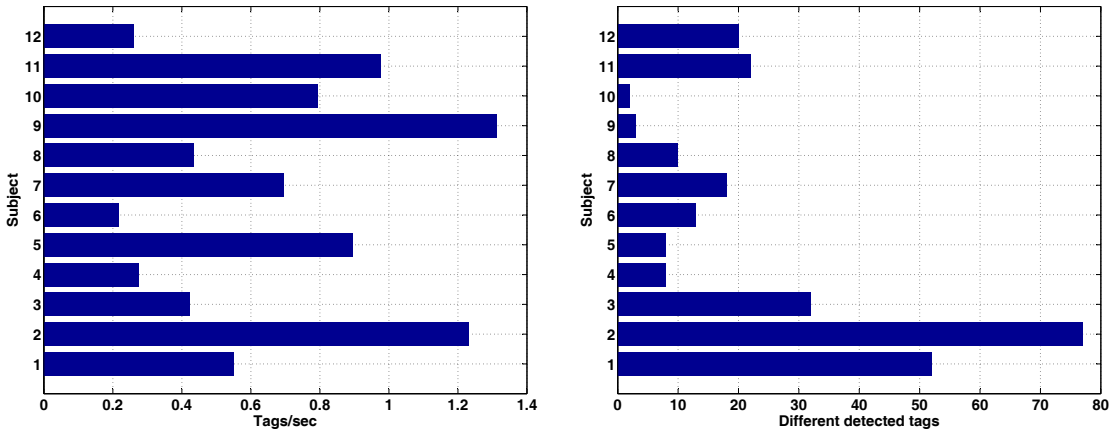
(a) Tag detections per activity (b) Different detected tags per activity

***Figure 4.5**: Per activity analysis*

assume that the way the tags had to be mounted to the mop, the broom, and the dusting cloths affects their readings. Another issue might be the object's material properties, especially for the metal handle of the broom that we used in the experiments. A smaller plastic hand broom was easier to detect, but the subjects used it only occasionally at the end of the brooming activity to gather the dust in a dust pan. Interestingly, the dish washing activity has relatively high tag detection score (0.85tags/sec) even though many tagged objects related to that activity (e.g. glasses and plates) are never detected. After inspection of the recorded RFID data, we noticed that most of the tag detections occurred during dish washing when two subjects used the gloves that were easily detected by a wrist-worn RFID reader. In cases when subjects did not use the gloves, there were almost no RFID readings for the dish washing activity.

The number of different detected tags per activities is shown in Figure 4.5(b). The bed making activity has the highest score (i.e. 42 tags) due to the fact that the objects used for that activity were tagged with multiple tags (pillow 18 tags, bed sheet 9 tags, blanket 9 tags, pillow cases 6 tags each - see Table 4.1 in Chapter 3). On the other hand, the dusting and setting the table activities in its nature include interaction with different objects, which is the reason why the more different tags were detected during these two activities (i.e. 18 tags and 17 tags, respectively) comparing to the others. The watering plants activity has the lowest score, i.e. only 2 different tags were detected for that activity. However, since we did not tag the plants, only 2 objects relevant for that activity were tagged (i.e. water spray and watering can). We also did not tag the clothes being ironed. Therefore, the ironing activity also has relatively low score (i.e. 8 tags).

We evaluated how much the number of tag detections (Figure 4.6(a)) and different detected tags (Figure 4.6(b)) varies among the subjects. As different subjects performed different sets of the activities, the scores between subjects vary significantly. Two subjects (subject 9 and subject 10) participated only in the ironing activity, and their tag detection scores are among the highest and the numbers of different detected tags are the lowest



(a) Tag detections per subject (b) Different detected tags per subject

***Figure 4.6**: Per subject analysis*

due to the high tag detection score and only a few different detected tags for that activity.

4.3.2 Activity Performance Diversity

As already stated in Section 3.1.1, we kept the scenario presented to the subjects very general. That resulted in a variety of ways the subjects performed the activities. The subjects had different interpretations of the activities. For example, two subjects used the vacuum cleaner not only to vacuum the floor, but also to clean the sofa. Surprisingly, five subjects did detailed vacuuming of the floor under the sofa. In two cases of the vacuuming activity, the subjects used their leg to switch the vacuum cleaner on and off, making it impossible for our sensors to detect those parts of the activity. Some of the typical actions that have to be done during an activity, such as pulling out the vacuum cleaner cable consisted of very different movements among the subjects. During the ironing activities, only one subject used the distilled water. None of the subjects placed the iron on the side handle of the ironing board. That significantly decreased the detection of the interactions with the ironing board, because the tags were placed close to the handle in hope that the subjects would use it during the ironing. Also, repetitive cleaning movements during activities such as brooming and vacuuming varied in intensity among different subjects. All these issues make the recognition task very challenging.

All our subjects wore the sensors on their dominant wrist. Still, during the ironing activity two subjects occasionally used their non-dominant hand for ironing some parts of the clothes that were easily reachable in that way. The same pattern occurred when one of the subjects was using both hands during dusting. Another problem is that in many situations, the dominant hand was occupied with another action, and the subjects had to interact with the necessary objects by using the non dominant hand. The RFID reader could not detect those events.

4.4 Approach

The main goal of our experiment is to study the combination of RFID and accelerometer sensing technology for ADL recognition. In order to do that, we first use the accelerometer and RFID tags separately, and afterwards we apply an integrated approach to overcome the shortcomings of both approaches. In the following, we describe all three approaches used.

4.4.1 Recognition Based on Acceleration Data

The 3D-acceleration data as recorded from the sensor is downsampled from 250Hz to 100Hz for our experiments. We compute the following features from the raw signal: *mean*, *variance*, *area under curve*, *energy*, *spectral entropy*, *pairwise correlation between the three axes*, the first ten *FFT coefficients* and *exponential FFT bands* [Lester *et al.* 2005]. Each feature is computed over a sliding window shifted in increments of 0.5sec. We evaluate the performance of the features both individually and in combination, and over different window lengths (0.5sec-128sec).

For classification of activities we evaluate three different approaches, namely *Naive Bayes* (see Section 3.2.1), *Hidden Markov Models* – HMMs [Rabiner 1989] (see Section 3.2.2) and *Joint Boosting* [Torralba *et al.* 2004] (see Section 3.2.3). In our experiments we use the unimodal Gaussian model for the Naive Bayes classifier.

4.4.2 Recognition Based on RFID Data

We associate all tagged objects with the activities in which they are typically involved. This process is done manually, but as we aim at tagging fewer objects, that should not be a major constraint for the implementation of our approach. With this object-activity mapping, each detected tag clearly indicates a candidate set of possible activities.

In Section 4.3 we have seen that the used dataset contains very few false positive readings, i.e. accidentally detected interactions with objects. Also, many interactions with objects are not detected, mostly due to the short range of the RFID reader's antenna. The number of tag detections highly varies among the activities. Figure 4.7 shows raw acceleration and RFID data for two activities, i.e. vacuuming (left column) and mopping (right column). As can be seen, tags were detected more often during vacuuming than during mopping. To overcome the problem of sparse tag detections, we use a sliding window over the detected RFID tags and classify each window based on the *weighted majority voting* scheme of the tag readings, i.e. mapped activity labels in that window. We shift the window in increments of 1sec and we evaluate the recognition performance for different window lengths (1sec-120sec). Additionally, the tags' votes are *weighted* proportionally to their relative position in the window, in order to avoid bias from the previous activity at activity transitions for longer window lengths.

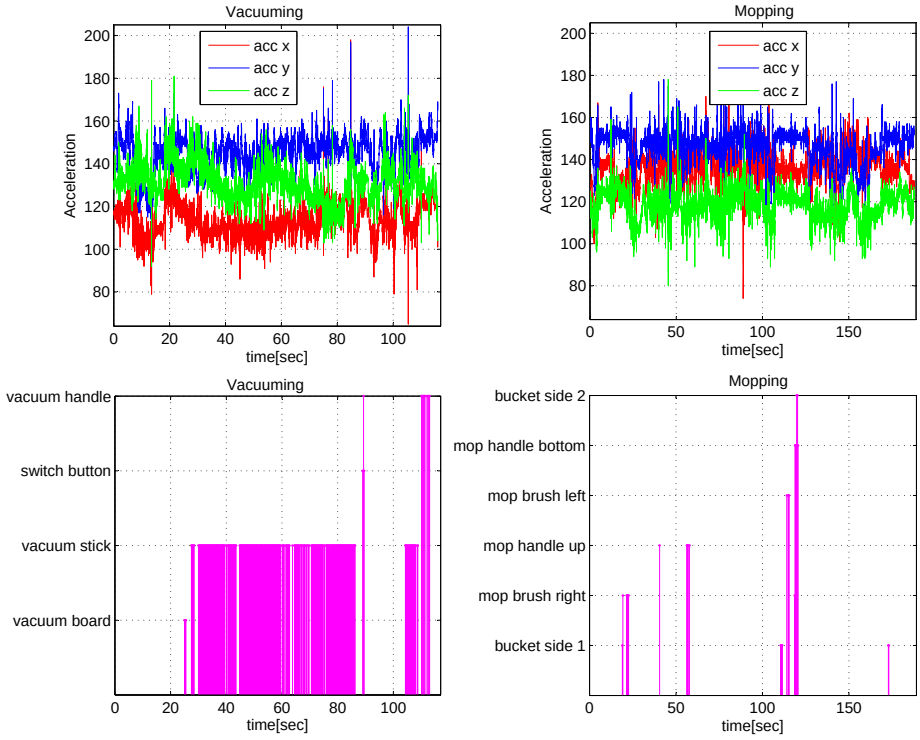


***Figure 4.7**: An example of raw acceleration and RFID tag data for two activities: vacuuming (left) and mopping (right).*

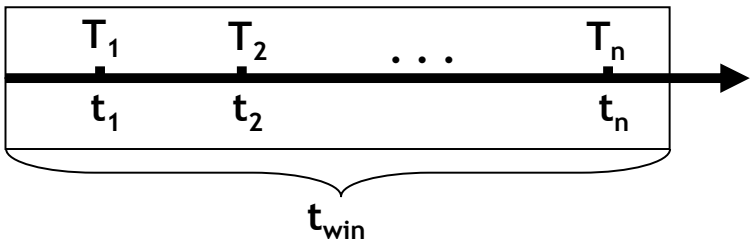


***Figure 4.8**: RFID sliding window approach based on weighted majority voting.*

Weighted Majority Voting. More formally, let C be the number of activities $\{A_1,...,A_C\}$. Figure 4.8 shows the RFID data observed in a sliding time window of length t_{win} (Figure 4.8). The tags $T_1,...,T_n$ are detected at times $t_1,...,t_n$. Each tag, T_j votes for a set of activities $S_j = \{A_{j1},...,A_{jk}\}$ it is involved in by the following rule:

$$v_{ij} = \begin{cases} 0, & A_i \notin S_j \\ \frac{1}{k}, & A_i \in S_j \end{cases} \tag{4.1}$$

where $i \in \{1,...,C\}$, and $j \in \{1,...,n\}$. Furthermore, each vote is weighted by the relative position of the detected tag T_j in a current sliding window:

$$w_j = \frac{t_j}{t_{win}},\ j \in \{1,...,n\} \tag{4.2}$$

The final votes for activities $A_i,\ i \in \{1,...,C\}$ are as follows:

$$V_i = \sum_{j=1}^{n} w_j v_{ij} \tag{4.3}$$

The current sliding window is classified as:

$$\hat{y} = \underset{i \in \{1,..,C\}}{\operatorname{argmax}} (V_i) \tag{4.4}$$

In case of the equal votes for multiple activities $W = \{A_{w1},...,A_{wm}\}$, the window is classified as a random activity from the set W of the winning activities:

$$\hat{y} = random(1,...,m) \tag{4.5}$$

In case of an empty sliding window when no tags is detected (i.e. $V_i = 0,\ i \in \{1,...,C\}$), the window is classified as an *unknown* activity $C+1$:

$$\hat{y} = C+1 \tag{4.6}$$

Activity	100%	50%	25%	12.5%
Bed making	42	21	11	6
Dusting	18	9	5	3
Setting the table	18	9	5	3
Cleaning windows	14	7	4	2
Mopping	8	4	2	1
Ironing	7	4	2	1
Vacuuming	5	3	2	1
Brooming	3	2	1	1
Dish washing	3	2	1	1
Watering plants	2	1	1	1

***Table 4.2**: Number of tags in different runs*

For the evaluation of the influence of the number of used tags on the recognition results, we use the following procedure (Table 4.2). In the first run, we use all deployed tags. Since we aim to tag as few objects as possible, we decrease the number of used tags in each run by half: We rank the tags for each activity based on the number of detections and then use the best 50% of these tags until in the last and fifth run we only have one tag per activity. Some of the activities include fewer objects than others, which is reflected in the dataset. E.g. activities such as *watering plants* and *brooming* require fewer interactions with objects. In some other activities, such as *washing dishes*, tagged objects are not detected, probably due to the absorption of the radio waves by water and metal.

4.4.3 Combining RFID and Accelerometer Sensing

The RFID reader provides accurate high-level information for the activity inference about the current user-object interaction, producing almost no false positive readings. Thus, for the combination of RFID and accelerometer sensing, we use the RFID recognition as a baseline method for the recognition of activities. In cases when we fail to recognize the activity based on RFID tags, we rely on the accelerometers' recognition. In principle, there are two different cases when the RFID approach fails.

In the first case, the majority of detected tags within a window is *shared* among several activities. Based on the RFID approach, the window is classified as one of the activities that share the tags in that window, each of those activities having the same probability (Equation 4.5). To decrease the classification errors, we resolve this ambiguity by using the acceleration classification. Let acceleration predictions within such an RFID time window be $\hat{y_1}^{acc},...,\hat{y_p}^{acc}$ with corresponding confidence scores $h_1,...,h_p$. For each activity $A_{wi},\ i \in \{1,...,m\}$ in W, we calculate its cumulative acceleration based likelihood as:

$$l_{wi} = \sum_{i=\hat{y_j}^{acc}} h_j \tag{4.7}$$

We classify the window as the activity which has the highest cumulative likelihood among the activities in the set W that share the detected tags in that window:

$$\hat{y} = \underset{i \in \{1,..,m\}}{\operatorname{argmax}} (l_{wi}) \tag{4.8}$$

In the second case, the RFID reader fails to detect any tags within a window. As we do not have any information about the current activity based on the RFID data, we classify the window as an unknown activity (Equation 4.6). We resolve the issue of gaps in RFID data by using again acceleration predictions for that RFID time window. We first calculate cumulative acceleration based likelihoods for each activity $A_1,...,A_C$ by accepting the acceleration based classifications only if their likelihood is above a certain threshold t_h:

$$l_i = \sum_{\substack{i=\hat{y_j}^{acc} \\ h_j > t_h}} h_j,\ i \in \{1,...,C\} \tag{4.9}$$

The window is classified as the activity with the highest cumulative acceleration based likelihood:

$$\hat{y} = \underset{i \in \{1,..,C\}}{\operatorname{argmax}} (l_i) \tag{4.10}$$

4.5 Results

In this section we present the experimental results for the three approaches described in Section 4.4. We evaluate our algorithms on the *Housekeeping* dataset presented in Section

3.1.1 by using standard metrics, namely precision, recall, and accuracy defined in Section 3.3.2. The ground truth for the sliding window used in all three approaches is the label of the last sample in the window. In order to examine the feasibility of person-independent activity recognition, we perform a 12-fold *leave-one-person-out* cross validation (see Section 3.3.1) on the data. All results are averaged over 12 cross validation runs.

4.5.1 Acceleration Results

In the following we report on our recognition results based on features computed from the wrist-mounted accelerometer alone. For the HMMs, we use the ergodic model and in addition to the window length, we vary the number of states (1-4), the number of Gaussians per state (1-4), and the observation sequence length (1-32). For Joint Boosting we vary the number of weak classifiers (100-200). In order to filter out occasional misclassifications, the output of all classifiers is smoothed with a majority filter.

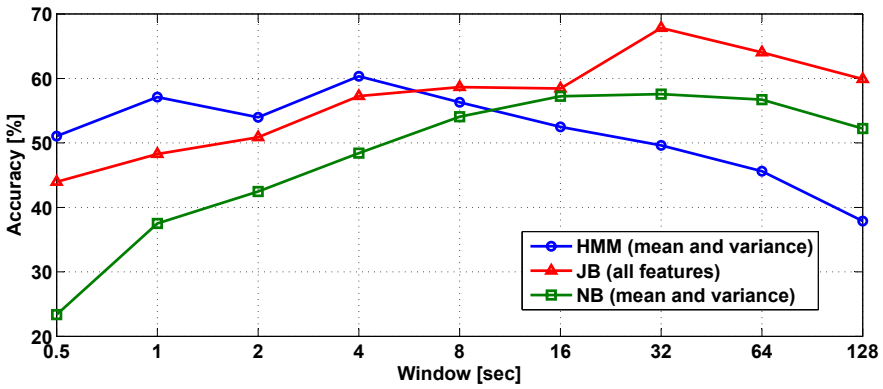


***Figure 4.9**: Classification based on data from the accelerometer. The plot shows the accuracy across different algorithms and window lengths.*

Figure 4.9 shows the results of our evaluation for different classifiers and window lengths. The figure shows the accuracy for the best parameter combinations. The overall best result of 68% accuracy is achieved with the Joint Boosting classifier when using all features and 200 weak classifiers. For Naive Bayes and HMMs, we found that using mean and variance of the signal as features works best for our set of activities. From Figure 4.9 one can observe that both Joint Boosting and Naive Bayes work best at relatively large window sizes of 32sec, while HMMs perform better at smaller window sizes of up to 4sec. One reason for this might be that the smaller windows preserve more of the temporal structure inherent in the data, which the HMMs are able to exploit. Table 4.3 shows the best HMMs' parameter combinations for different window lengths. Typically, the best results are achieved with 4 states. Modeling the targeted set of complex housekeeping activities might benefit from using a larger number of the HMMs' states. However, in that case time required for training the models increases significantly.

Window length [sec]	Number of states	Number of mixtures	Observation sequence
0.5	4	4	32
1	4	2	4
2	4	1	1
4	4	1	1
8	1	4	1
16	2	1	4
32	1	2	32
64	4	1	32
128	1	1	4

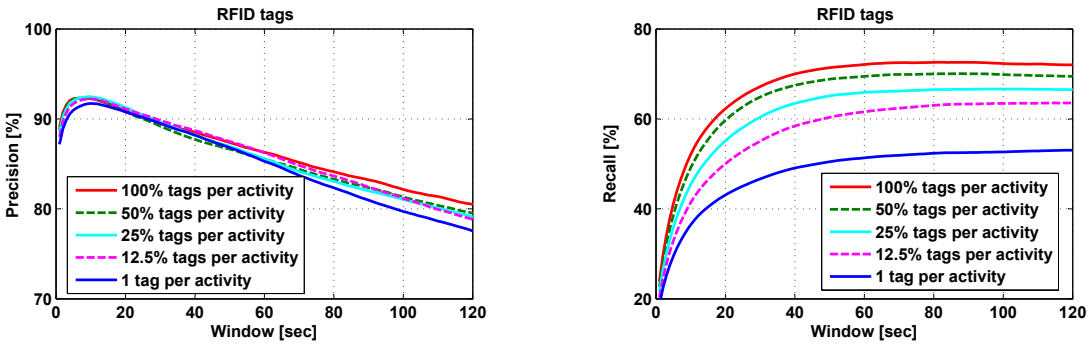
***Table 4.3**: HMMs parameters for different window lengths.*

The seemingly low accuracy of slightly below 70% should be seen in the light that there were several factors making this recognition task more challenging than others reported in the literature: First, the use of only a single 3D accelerometer, and second the fact that we train and test the system on different users, some of which performed the same activity in distinctly different ways and sometimes with different hands. Third, we did not edit the recordings e.g. by cutting out only the part during which the user actually ironed, but we included the entire activity from setup (e.g. assembling the ironing board) to teardown (e.g. stowing away the ironing board).

4.5.2 RFID Results

In the following we report on the recognition results based on the RFID tags only. Figure 4.10(a) and Figure 4.10(b) show how overall precision and recall change with different window lengths. One can observe that our approach performs best in terms of precision for very short windows. On the other hand, recall is higher for longer windows. That is due to the fact that with longer windows we propagate the labels to the regions where tags were not detected, so we have fewer false negatives. At the same time, longer windows increase the number of false positives, because of the tags' bias from the previous activity at the transitions between activities. When using all deployed tags, the highest results for precision lie slightly above 92% when using windows of 7sec. Recall reaches its maximum of 72% for windows of 82sec. Since recall dramatically increases when we increase the window length from 1sec to 40sec, and afterwards only a slight improvement is achieved, we propose to use 40sec windows as an optimal window length in this setting. Precision in that case still remains high (89%/70% precision/recall).

Figure 4.10(a) and Figure 4.10(b) also show the effect of using different numbers of tags. As can be seen from the plots, by decreasing the number of tags, recall decreases as well, but surprisingly precision does not change much. In some cases precision even increases with fewer tags, because some of the tags shared among the activities are discarded from the dataset in that way. When we use only one tag per activity, we choose the tag that was detected most often during the execution of each activity. That way, we



(a) Precision (b) Recall

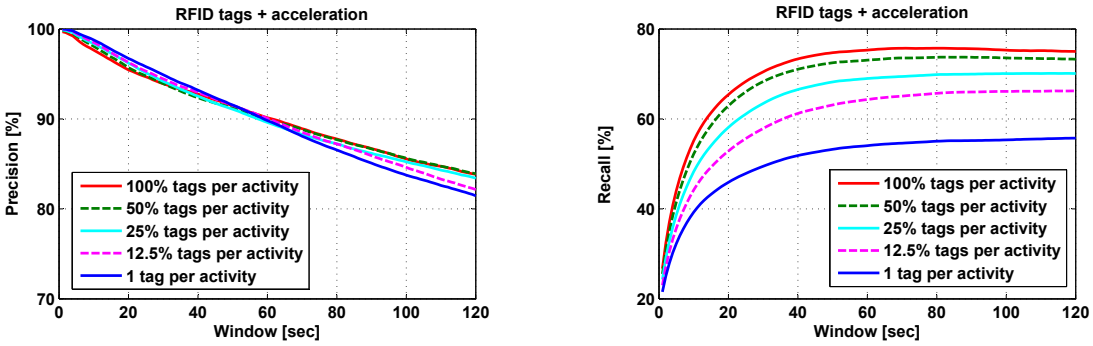
***Figure 4.10**: Overall precision and recall for different window lengths and different number of used tags in case of the recognition based on the RFID tags.*

define the key objects per activity (Table 4.4). Since the run with 12.5% of the most detected tags performs overall better than the run when we use only one tag per activity, we add three more objects to the set of key objects. That way, we also avoid the gloves as a single key object for *window cleaning*, since they are shared among that activity and *washing dishes*.

Activity	Key object	Additional key object
Bed making	Pillow	Pillow case
Dusting	Dusting cloth	Duster's box
Setting the table	Glass	Cupboard
Cleaning windows	Gloves	Window cleaning liquid
Mopping	Mop	-
Ironing	Iron	-
Vacuuming	Vacuum cleaner	-
Brooming	Small plastic broom	-
Dish washing	Gloves	-
Watering plants	Water spray	-

***Table 4.4**: Key objects for activities*

We tagged most of the objects with multiple tags to find the best placement for the tags. Here, we suggest the best placement of the tags for some of the key objects. For many objects (e.g. vacuum cleaner, mop, broom, iron) the tags placed on the handle of the object were detected more often than the other tags attached to the same object. That is due to the very short distance between the object handle and the RFID reader during the performed activities. For example, the tag on the handle of the vacuum cleaner was detected more often than the other 8 tags attached to it. For other objects the best placement is at the place where users usually grab the object (e.g. corner of the pillow) or at the place where users spend considerable time during the execution of the activity (e.g. buttons on the pillow case). For some objects (e.g. cupboard, window cleaning liquid, and dusters' box) the best placement depends on whether the subject is left or right-handed. We tagged the cupboard with 2 tags, close to the opening handle. The tag conveniently



(a) Precision (b) Recall

***Figure 4.11**: Overall precision and recall for different window lengths and different number of used tags in case of the recognition based on the RFID tags and acceleration for shared tags.*

placed to be detected by the RFID reader of the right handed persons was detected more often than the tag conveniently placed for left-handed persons, because we had only 3 left handed subjects. However, the tag on the window cleaning liquid placed close to the RFID reader of the left handed persons was detected more often than the tag placed close to the RFID reader of the right handed persons. That is due to the fact that most of the subjects used their non-dominant hand to grab the window cleaning liquid, since their dominant hand was busy with opening the window.

Overall, our experimental results show that a satisfactory trade-off between precision and recall can be achieved with appropriate window lengths. The reduced number of tags does not decrease the recognition results significantly and the key objects for activities are defined. Finally, the best placement for the tags highly depends on the person, as well as on the activity.

4.5.3 Combining RFID and Accelerometer Sensing Results

In the following we report on the recognition results based on the combination of RFID tags and acceleration. As shown in Section 4.4.3, we augment the RFID classification with acceleration recognition scores in two cases: 1) when detected tags are shared among the activities and 2) when interactions with objects are not detected. For the combination of RFID and acceleration classification, we use the parameters that yield the best results for the classification of acceleration data (i.e. Joint Boosting, all features over windows of 32sec).

We present the results of resolving tag ambiguities by means of acceleration classification in Figure 4.11(a) and Figure 4.11(b). We have only 4 types of objects (tagged with 16 tags) that are shared among 5 activities in the dataset. Still, compared to the results when we use all tags for the classification based on RFID tags only (Figure 4.10(a) and

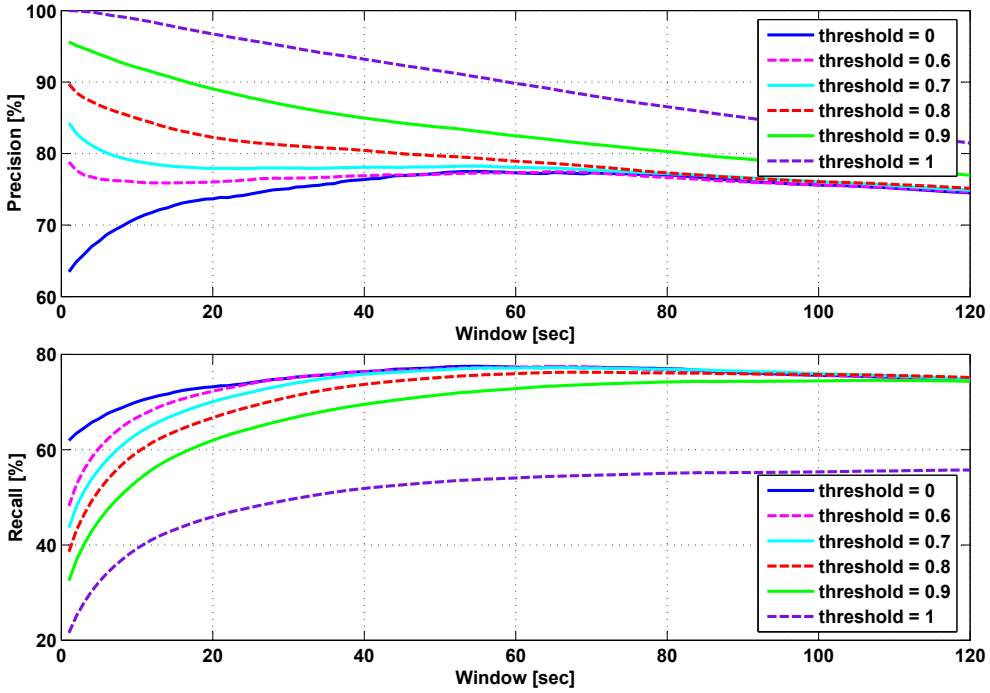


***Figure 4.12**: Overall precision and recall for different window lengths and different likelihood thresholds in case of 1 used tag per activity.*

Figure 4.10(b)), there is a clear tendency of about 3% improvement in recall. The precision increases, especially for shorter windows (from 10% increase for windows of 1sec to 6% increase for windows of 7sec, when the classification based on the RFID tags reaches its maximum). For larger windows, the gain in precision is lower but still noticeable (for windows of 40sec, the increase is 4%, and for the largest windows of 120sec, there is still increase of 3%). This decrease of improvement for larger windows is due to the fact that in larger windows, we typically have not only the shared tags, but also additional tags that resolve the tag ambiguities already on the RFID classification level.

The results of additional filling in of gaps where no RFID tags are sensed by using the acceleration classification are shown in Figure 4.12. Here, we present the results for the run when we use only one tag per activity. We vary the threshold between 0 (when all acceleration based classifications are accepted) and 1 (when all acceleration based classifications are rejected, which brings us to the previous case of using the acceleration predictions for shared tags only). From the plot one can observe that recall increases with the number of accepted acceleration based classifications. However, the more accepted acceleration based classifications we have, the more precision decreases. This is due to the fact that the recognition of higher level activities such as housekeeping is difficult using only one accelerometer placed at the dominant wrist of a user. This trade-off between precision and recall has to be taken into account based on the specific application requirements.

In the extreme case, when the threshold is 0, there is no unknown sample in the test

data, which means that overall precision and recall become the same. For shorter windows the increase of recall is between 40% for windows of 1sec and 33% for windows of 7sec. At the same time, we observe a significant decrease of precision (from 100% to 63% for windows of 1sec, and from 99% to 69% for windows of 7sec). For larger windows, the trade-off between precision and recall is better, since we have higher increase of recall compared to the decrease of precision. For example, for window length of 40sec, the recall increases by 24% and precision decreases by 17%. For the largest window of 120sec, the decrease of precision is almost three times lower than the increase of recall, i.e. precision decreases by 7% and recall increases by 19%. This is most likely due to the fact that the probability that there is no detected tag is lower for larger windows than for shorter windows. Therefore, the shorter windows need to rely more often on acceleration based classification which decreases the precision.

Classification results are slightly higher in the run when we use 100% tags, but the overall improvement in precision/recall is higher in the run when we use only one tag per activity. Thus, we can achieve good recognition results with only a few RFID tags when combining them with accelerometer sensing.

Discussion. The activity recognition scheme based on the combination of RFID and accelerometer sensing yields better recognition scores than either sensing technology alone. However, in order to augment the manual assessment of ADLs the proposed approach needs to overcome a few limitations.

The main issue in the RFID part of the system is a significant number of false negatives, i.e. tags were not detected even though the subjects were interacting with the tagged objects. This is due to the short range of the RFID reader but also due to the usage of the non-dominant hand in some activities. For example, during the *ironing* activity two subjects occasionally used their non-dominant hand for ironing some parts of the clothes that were easily reachable in that way. Also, in four cases of *cleaning windows*, the dominant hand of the subjects was occupied with cleaning utensils and subjects had to open the window with the non-dominant hand. Therefore, an additional bracelet on the non-dominant hand might improve the number of detected tags, with a risk of a lower user acceptance of the system. Another issue encountered during the experiment is tag ambiguities, i.e. an object is used in more than one activity. We aim to overcome these problems by relying on acceleration classification.

For the activity classification based on the accelerometers, we apply state-of-the-art algorithms, but the recognition scores for the acceleration part of our system still encounters issues most likely because of the following reasons.

First, in our experiment, we aim at person-independent training with 12 subjects who performed activities in very different ways. As already mentioned, two subjects vacuumed not only the floor but also the sofa. The subjects also had different strategies for *dish washing*. Three subjects did the washing by repetitive scrubbing and rinsing of each dish and the other subjects first scrubbed and then rinsed all dishes.

Second, we had three left-handed subjects and even by visual inspection of the acceleration data it can be clearly seen that the range of their data is not the same as for the right-handed subjects. One possible solution to this problem would be to train and test the algorithms only on right-handed or left-handed people.

Third, we use only a single accelerometer worn on the wrist of the user which causes lower recognition scores than usually presented in the literature. Some of the specific movements during the execution of the activities could not be inferred. As already stated, two subjects were turning the vacuum cleaner on and off with their foot. Additional accelerometers would increase the accuracy of the system, but again with a risk of lower user acceptance.

Fourth, we did not divide the activities into phases since we wanted to avoid scripted activity stages, as well as their temporal modeling, training and labeling. After comparing the recognition results with the ground truth and the video recordings, we found that the acceleration classifiers often fail to recognize parts of the activities that do not include discriminative movements typical for that specific activity. For example, during the *ironing* activity, parts when users were really ironing were correctly recognized, but parts when users were finishing ironing of one piece of clothing and preparing the next piece of clothing were usually misclassified. Also, different users performed beginning and ending of the activities differently, which introduced additional misclassifications.

4.6 Conclusion

The main goal of this chapter was to demonstrate the feasibility of combining RFID and accelerometer sensing for ADL/IADL recognition. We conducted an evaluation of our algorithms' performance on 10 housekeeping activities, executed by 12 subjects. Detailed analysis of the algorithms' parameters indicates the optimal window lengths and features, which are 40sec window for RFID based recognition and 32sec window and combination of all acceleration features for the Joint Boosting approach. The results show that combined recognition helps in cases of tag ambiguities, i.e. when tagged objects are being shared among the activities, as well as in periods when the RFID reader can not detect interactions with objects due to its short range.

We aim to decrease the number of tagged objects and accelerometers worn by users, while keeping satisfactory recognition results when combining the two sensor modalities. By using different numbers of tags in the dataset, we explored how the number of tags influence the recognition. The results indicate that a decreased number of tags does not significantly change the precision of our system. In some cases, by decreasing the number of tags, tag ambiguities disappear from the dataset, increasing the precision. This supports the assumption that the tags should be strategically placed on the key objects.

In order to make the deployment of the activity recognition system in home environments feasible, in the next chapter we further explore the potential of multi-sensor

approaches for reducing the level of supervision in activity recognition. This would allow investigation of the algorithms' generalization capabilities on larger datasets recorded over longer periods of time.

5

Towards Less Supervision Based on Complementary Sensors

In the previous chapter we have seen that a multi-sensor approach can accurately recognize activities by incorporating knowledge about different activity characteristics. In this chapter we continue our work in the multi-sensor direction, by exploring its feasibility for reducing the level of supervision in activity recognition. As stated in Chapter 1, the generation of labeled training data is not only tedious and error prone but also limits the applicability and scalability of today's activity recognition approaches. Thus, in this chapter we systematically analyze and compare two different techniques to significantly reduce the required amount of labeled training data, namely *semi-supervised* learning and *active* learning. In our experiments we employ two complementary sensors for inferring body-motion and location, which are important characteristics of daily activities. The experimental results suggest that both techniques obtain similar and sometimes even better performance than standard supervised techniques, while using only a limited amount of labeled training data.

5.1 Introduction

The primary goal of this chapter is to explore and compare two different types of techniques that require far less labeled training data than traditional supervised techniques. First, we apply and analyze the merits of two of the most fundamental semi-supervised learning techniques, namely *self-training* and *co-training*. Typically, in semi-supervised settings, it is assumed that in addition to the small set of labeled training data there is also a substantial amount of unlabeled training data available. This allows reducing the effort of supervision to a minimum, while still preserving competitive recognition performance. And second, we also explore another way to reduce the required amount of labeled training data. This second approach is based on active learning [Muslea *et al.* 2000] with the explicit goal to focus labeling effort on the most profitable, e.g. informative, instances of activities.

The main contributions of the chapter are as follows. First, we present a comparative evaluation of the applicability of self-training [Chapelle *et al.* 2006] and co-training [Blum and Mitchell 1998] for data from motion sensors. Unlike in [Guan *et al.* 2007], where an ensemble method based on one set of features has been proposed, we show that it is possible to apply co-training for recognition of activities when using two independent complementary sources of information, namely on-body accelerometers and infra-red motion sensors. Second, we suggest two functions to actively probe users for labels that enable active learning. The wrapper nature of the proposed semi-supervised algorithms and active sampling functions makes them independent of both classifiers and sensor modalities being used. Additionally, their low computational costs are very beneficial for enabling real-world scenarios. Third, we enhance the efficiency of the proposed activity recognition system by utilizing a multi-class boosting procedure, namely Joint Boosting [Torralba *et al.* 2004] introduced in Section 3.2.3. Additionally, the typical researchers' bias on the evaluation is avoided by using a publicly available dataset *PLCouple1* (see Section 3.1). By using only a limited amount of labeled training data, we achieve performance comparable to and sometimes even better than fully supervised learning approaches on a challenging and realistic dataset.

The rest of the chapter is organized as follows. In Section 5.2 we describe our experimental setup. Section 5.3 presents the initial supervised analysis of the dataset followed by our semi-supervised and active learning approaches in Section 5.4 and Section 5.5, respectively. Finally, Section 5.6 concludes the chapter.

5.2 Experimental Setup

In the field of activity recognition, the state-of-the-art has advanced significantly in recent years and a wide range of sophisticated approaches and sensors has been developed. We have argued in Chapter 1 that an important drawback of the majority of current activity recognition systems is the lack of a standardized evaluation procedure. Thus, in this and the following chapters we follow a different approach by using the PLCouple1 [Logan *et al.* 2007] dataset provided by a second party.

In the experiments we use acceleration and infra-red data. Unlike in [Logan *et al.* 2007] where the mean value of the acceleration signal and binary occurrences of the infra-red readings were used as features, we extract the following features to exploit the full richness of information in the data: 1) From the raw acceleration signal we compute *mean*, *variance*, *energy*, *spectral entropy*, *area under curve*, *pairwise correlation between the three axes*, and the first ten *FFT coefficients*, which sums up to 48 features per acceleration sensor channel. 2) For each of the ten infra-red sensors we calculate the number of their activations as features. As in [Logan *et al.* 2007], each feature is computed over a sliding window of 30 seconds shifted in increments of 15 seconds. We experimented with different window lengths as well, but that did not significantly change performance.

In [Logan *et al.* 2007] movement data measured by two accelerometers, worn on the

dominant wrist and on the dominant hip, were used. As the dataset includes data from a third accelerometer, worn on the non-dominant thigh, we perform the experiments with both two and three accelerometers since the addition of sensors often improves recognition performance.

As suggested in [Logan *et al.* 2007], we use 9-fold leave-one-day-out cross validation (see Section 3.3.1) on the data to avoid overfitting. In each cross validation round of supervised learning, we train the algorithms on 8 days of data. In case of semi-supervised and active learning, only a subset of 2 days of data is used as an initial labeled training set. The algorithms are always tested on the left out day's data.

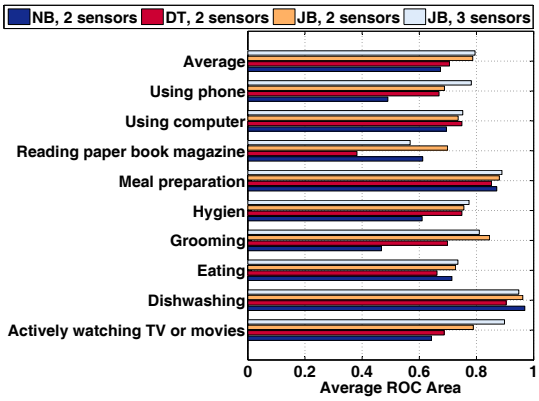
5.3 Supervised Approach

As we use the publicly available subset of the PLCouple1 dataset, we first reproduce the experiments from [Logan *et al.* 2007] based on two supervised machine learning algorithms: Naive Bayes (see Section 3.2.1) and Decision Trees (see Section 3.2.4). Additionally, we compare their performance to the Joint Boosting classifier [Torralba *et al.* 2004] (see Section 3.2.3). These results are used as a baseline for comparison with semi-supervised and active learning approaches in Section 5.4 and Section 5.5, respectively.

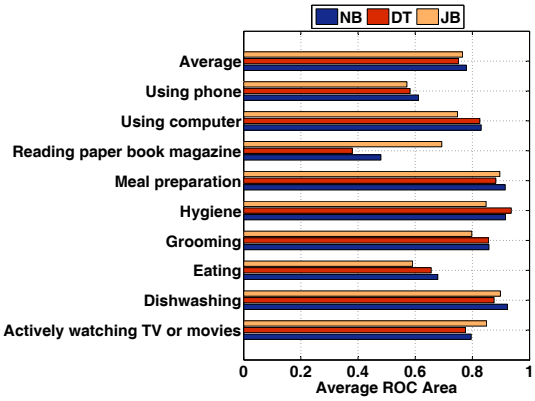
Since the dataset contains partly overlapping activities that are not mutually exclusive, here we use, as in [Logan *et al.* 2007], the area under the Receiver Operating Characteristic (ROC) curve (see Section 3.3.2), averaged over 9 cross validation rounds, as a figure of merit. For Naive Bayes and Decision Trees we apply a "one vs. rest" approach, as in [Logan *et al.* 2007], by using a binary classifier for each activity. The main drawback of that approach is that it does not deal well with highly unbalanced datasets. The overall duration of activities in the dataset strongly varies among the activities, reflecting the natural distribution of activities in real life. Thus, the balancing of the training set had to be done, as in [Logan *et al.* 2007], by uniformly sampling the examples from the negative class to match the number of examples in the positive class, i.e. activity of interest. Interestingly, Joint Boosting, being a multi-class classifier, lends itself to joint training on all classes by finding features that can be shared across the classes. As a consequence, it is able to deal properly with multi-label data of overlapping activities (i.e. activities that were performed in parallel which resulted in multiple labels for a single sample). We transformed multi-label samples to single-label samples as follows: Each multi-label sample consisting of n labels is replicated n times, and the i-th copy is assigned the i-th label. During classification we accept all classes with classification scores higher than a threshold.

5.3.1 Results

In the following we report the recognition results based on the supervised algorithms. We experimented with both binary features and the number of the activations of infra-red



(a) Per-activity and average results for acceleration data



(b) Per-activity and average results for infra-red data

Figure 5.1*: Leave-one-day-out cross validation results for supervised classifiers (Naive Bayes - NB, Decision Trees - DT, and Joint Boosting - JB).*

sensors. Here, we only report the best results per classifier, i.e. performance of Naive Bayes and Decision Trees for binary features and performance of Joint Boosting when using the number of activations as a feature. We perform the experiments with different numbers of Joint Boosting rounds. The best performance is achieved after 50 iterations for acceleration data and after 10 iterations for infra-red data. Since the acceleration feature vector has 144 components it requires more boosting rounds to find the best features to be shared among the activities. The infra-red feature vector has only 10 components and weak learners from additional rounds could not improve performance.

Figure 5.1(a) and Figure 5.1(b) show results per activity and average recognition performance for acceleration and infra-red sensors, respectively. A few trends stand out. First, one can observe that Joint Boosting yields better results for 7 out of 9 activities when using acceleration data only. On average, Joint Boosting improves the results by 11.3% compared to Naive Bayes and by 8.2% compared to the Decision Trees classifier. Second, the addition of the third accelerometer does not improve the results significantly, presumably because the placement of the sensor at the non-dominant thigh is not discriminative for the majority of the activities studied. Third, Naive Bayes on average performs slightly better for infra-red sensors. As stated in [Logan *et al.* 2007], the presence of a second subject in the apartment whose activities were not annotated introduced noise in the infra-red sensor data. Thus, Naive Bayes, as a generative model, is able to deal better with the noisy data compared to the Joint Boosting and Decision Trees classifiers. Even though, we use only the publicly available subset of the PLCouple1 dataset, the Decision Trees results are nearly the same as reported in [Logan *et al.* 2007].

As previously mentioned, the dataset contains a certain amount of overlapping activities. The multi-label data constitutes about 10% of the whole dataset. Table 5.1 summarizes the classification results of the Joint Boosting classifier when leaving out the multi-label part of the dataset. The results are consistent with the multi-label case (i.e.

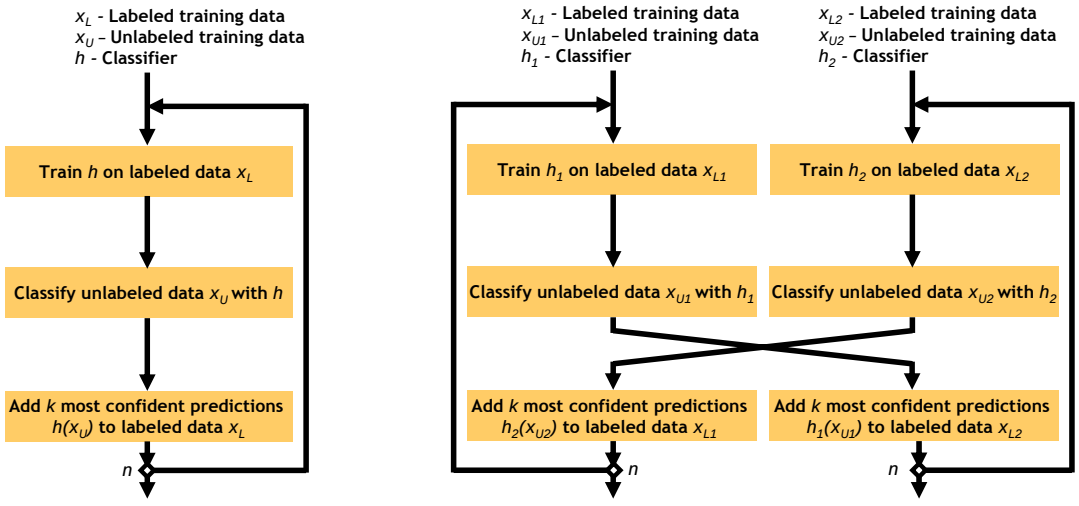
Sensor	Accuracy	Average ROC Area
Acceleration	53.6%	79.3%
Infra-red	41.6%	68.6%

***Table 5.1**: Leave-one-day-out cross validation results for Joint Boosting classifier on single-label subset of the dataset.*

Joint Boosting again performs better on acceleration data). Additionally, the table shows accuracy (defined in Section 3.3.2) of the classification. One can observe that accuracy is relatively low (53.6% for acceleration data and 41.6% for infra-red data), but that should be seen in the light of realism of the used dataset which additionally includes many other activities that were considered as an unknown class during the classification procedure. In order to thoroughly explore the potential of semi-supervised and active learning in activity recognition we decided to use a clean dataset (i.e. without multi-label samples) in the remainder of the chapter. The results in Table 5.1 are used as a baseline for comparison with semi-supervised and active learning approaches. As a figure of merit we use accuracy, which we consider more intuitive and which is more often used than the area under the ROC curve in the field of activity recognition.

5.4 Semi-Supervised Approaches

In this section we introduce the two semi-supervised approaches, *self-training* and *co-training*, which we use in our experiments for learning from both labeled and unlabeled training data.



(a) Self-training (b) Co-training

***Figure 5.2**: Semi-supervised algorithms.*

Self-training [Chapelle *et al.* 2006] is a wrapper-algorithm that repeatedly uses a supervised learning method in the following manner (Figure 5.2(a)). A supervised classifier h is first trained with a small amount of labeled data x_L. The classifier is then used to classify the unlabeled data x_U. In each iteration i, $i \in \{1,...,n\}$, a part of the unlabeled data is labeled according to a current decision function. Typically, *k most confident* predictions $h(x_U)$ are added to the labeled training set x_L. The classifier is then re-trained and the self-training procedure is repeated.

Co-training [Blum and Mitchell 1998] (Figure 5.2(b)) follows the iterative training procedure of self-training. At the same time, it aims to improve self-training by augmenting the training process with an additional source of information. Thus, we initially use acceleration and infra-red feature sets x_{L1} and x_{L2} for training two separate classifiers h_1 and h_2. Classifiers then teach one another by augmenting each other's labeled training sets x_{L1} and x_{L2} with their *k most confident* predictions $h_2(x_{U2})$ and $h_1(x_{U1})$, respectively. The classifiers are then re-trained with the refined labeled training sets and the process is iteratively repeated n times. Co-training is based on the two assumptions that are fulfilled in our multi-sensor approach. First, it assumes that features can be split into two disjoint sets that are sufficient for learning in the supervised setting so that one can trust the predictions based on both sets. Second, the two sets of features need to be independent given the class, so that one classifier's high confident data points are independent and identically distributed samples for the other classifier.

In [Guan *et al.* 2007] it has been argued that co-training is not applicable to activity recognition due to the strong independence assumption. In this chapter, we show that co-training is an excellent method for activity recognition approaches that aim at improving recognition results by fusing different sensor modalities. In the following experiments we use acceleration and infra-red data for co-training and compare its performance with self-training. Since Joint Boosting shows superiority compared to the Naive Bayes and Decision Trees classifiers, we use it as the supervised part of the self-training and co-training procedure.

The experiments are designed to investigate the trade-off between labeling efforts and recognition performance. The goal of the experiments is to decrease the amount of necessary labeled training data to a minimum. For that purpose, we use the following evaluation procedure (Figure 5.3). Leave-one-day-out cross validation is again performed by using one day of data for testing and the remaining eight days of data for training. The distribution of activities varies significantly for different days. Since we want to find the lower boundary for the size of labeled training data we use a minimum amount of data to have at least one sample for each of the activities of interest. In case of the used PLCouple1 dataset, that means that we can use six days of data as unlabeled training set and the remaining two days of data as an initial set for subsampling to get the reduced set of labeled training data. The experiments consist of five different configurations in which we gradually decrease the amount of labeled training data. These five configurations are constructed based on randomly sampled 50%, 25%, 10%, 5%, and finally only 1% of data from the selected two days. In each cross-validation round another two days of data are used for subsampling of labeled training set. As the amount of annotated data per

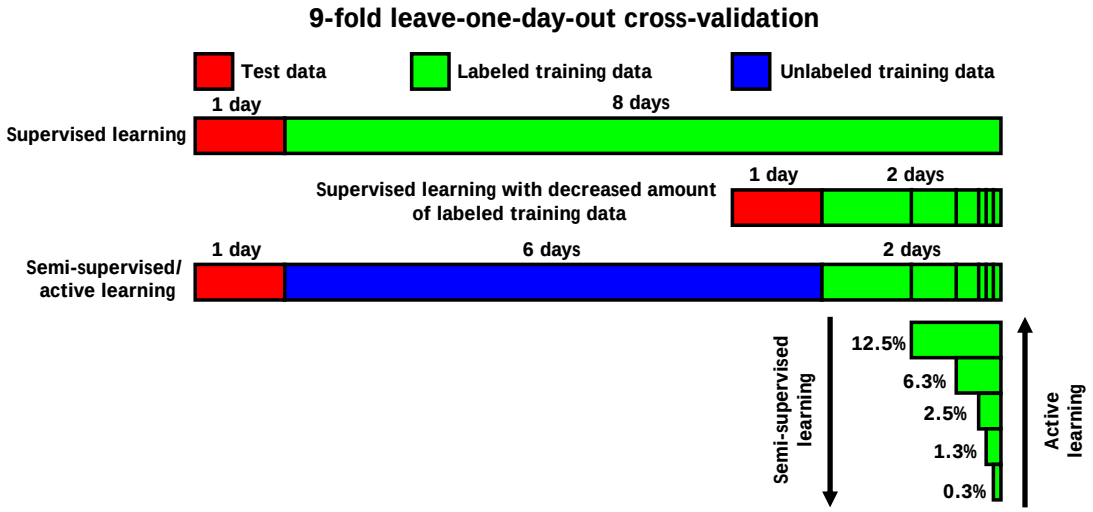


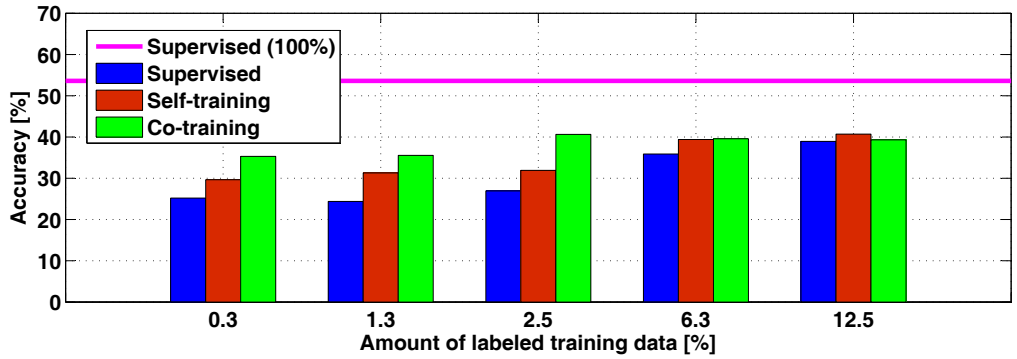
***Figure 5.3**: Evaluation procedure used in the experiments. In the semi-supervised experiments the amount of labeled training data is decreased by random subsampling to 12.5%, 6.5%, 2.5%, 1.3%, and 0.3%. In the active learning experiments, we start with 0.3% labeled training data and increase the amount of labeled training data to 1.3%, 2.5%, 6.5%, and 12.5% by active sampling functions. These two approaches are compared with the supervised learning approaches.*

day varies, these five configurations on average sums up to 12.5%, 6.3%, 2.5%, 1.3%, and 0.3% of the complete set of labeled and unlabeled training data. In order to thoroughly analyze the classifiers' performance we perform multiple random subsampling rounds. The reported results are averaged over 9 cross-validation and 5 random subsampling rounds. We compare the performance of the semi-supervised algorithms with the supervised approaches when using all training data as labeled and when using the reduced amounts of labeled training data (i.e. 12.5%, 6.3%, 2.5%, 1.3%, and 0.3%).

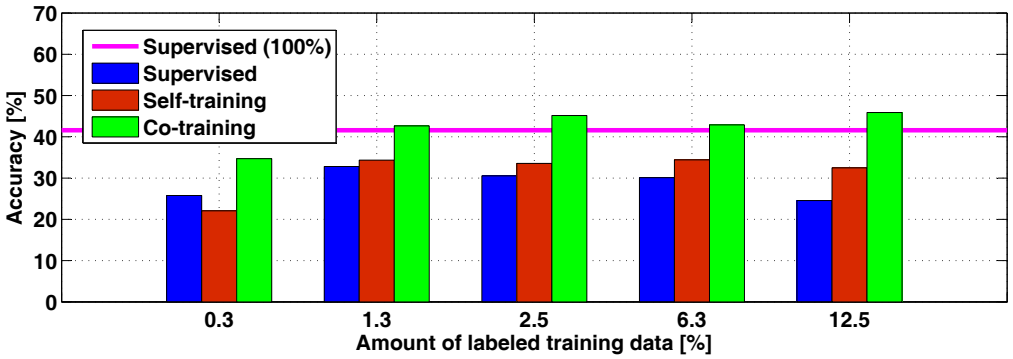
5.4.1 Results

An important parameter of the self-training and co-training algorithms is the number of iterations. By conducting experiments with different numbers of iterations we observed that by performing more than 100 iterations the newly labeled samples do not contribute any additional discriminative information, and at a certain point the labeling accuracy even starts to decrease. For comparison of self-training and co-training, in the following, we report on the average recognition accuracy achieved after 100 iterations.

We also observed that for our multi-class problem it is crucial to maintain the underlying distribution of activities. In each iteration we accept the 50 most confident predictions, but the number of accepted samples per activity needs to be matched to the initial distribution of activities in the labeled training set. We performed experiments with fewer



(a) Leave-one-day-out cross validation results based on acceleration data



(b) Leave-one-day-out cross validation results based on infra-red data

***Figure 5.4**: Comparative performance of self-training, co-training and supervised learning for different amounts of labeled training data.*

accepted samples per iteration, but in that case the learning phase is slower, because more iterations are required to achieve high performance. Additionally, in order to get more representative samples for the labeling process, as suggested in [Blum and Mitchell 1998], we carried out random sampling of unlabeled training data and performed the labeling on that subset of data. This, however, did not improve the results.

Figure 5.4(a) and Figure 5.4(b) show the classification accuracy of self-training (red bars) and co-training (green bars) when using different amounts of labeled training data for acceleration and infra-red sensors, respectively. The plots also show the comparison to the supervised approach (blue bars) when using the same decreased number of labeled training data, as well as the expected upper boundary (pink line) when using 100% of training data for supervised learning.

From the plots one can clearly observe a superiority of co-training compared to self-training, e.g. when using 2.5% labeled training data the performance of co-training is 12% higher than the performance of self-training on infra-red data. For acceleration data, accuracy increases by 9% when using the same amount of labeled training data, i.e. 2.5%. The performance of self-training on both sensor modalities, i.e. acceleration and infra-red does not differ significantly. For acceleration data there is a consistent improvement com-

Amount of labeled training data	100%	12.5%	6.3%	2.5%	1.3%	0.3%
Number of labels	9613	1203	604	244	124	29

***Table 5.2**: Average number of labels used for different experiment configurations.*

pared to the supervised approach with the same reduced amount of labeled training data. For infra-red data, after self-training the performance is sometimes degraded (when using 0.3% labeled training data), which highly depends on the quality of the initial labeled training subset of data.

The experiments in Section 5.3 show that Joint Boosting performs better on acceleration data than on infra-red data. Therefore, the full strength of co-training is clear when looking at the benefit that infra-red data gains from co-training. The performance is boosted by more accurate acceleration predictions during co-training. In most of the configurations, it outperforms even the supervised approach when using 100% labeled training data. For the configuration using 2.5% labeled training data, the performance of co-training is 4% higher than in the supervised case of 100% labeled training data. Co-training of acceleration data never achieves the performance of the supervised case of 100% labeled training data, but the strength of the algorithm is still visible compared to the supervised case when using the same reduced amount of labeled training data as for co-training. In the case of 2.5% labeled training data, the increase of performance is 3% for self-training and 14.6% for co-training. Surprisingly, by using more labeled training data performance of co-training starts to decrease, presumably due to the noise in infra-red data that is more inherent in larger random subsets of data.

All the above mentioned results clearly show the potential of semi-supervised approaches to minimize the labeling efforts. As can be observed from Table 5.2, the number of labels averaged over 9 cross validation rounds is extremely reduced compared to the average of 9613 labels when using 100% labeled training data for the supervised approach presented in Section 5.3. In the configuration when we use 2.5% labeled training data, as can be seen from Figure 5.4(a) and Figure 5.4(b), the achieved results are impressive, considering that only 244 labels are used. In that case, 6 activity models are learned with less than 5 labels per activity. When further decreasing the number of labeled training samples, some of the activities are learned from a single label. In the extreme case, when using 0.3% labeled training data, i.e. only 29 labels, 6 out of 9 activities are learned from a single labeled sample per activity. In that case the achieved performance is relatively low due to the very few labels, but by carefully choosing the data to be labeled the performance can still be significantly improved. Therefore, in the next section we utilize active learning to train activity models.

5.5 Active Learning Approach

Active learning aims at detecting the most informative unlabeled samples and queries a user to label them. In the context of activity recognition, one can legitimately imagine an online algorithm, similar to the stream-based setting in [Kapoor and Horvitz 2008], that asks the user to annotate his current activity when it is considered necessary for improving the recognition performance.

We employ a multi-sensor approach for active learning to select important samples to be labeled. The approach is based on a pool-based setting, i.e. we use a small set of labeled data and a large set of unlabeled data for training. The active learning algorithm searches for samples from the unlabeled training data to be labeled by a user. Two *active sampling* functions are evaluated here. The first function is based on the assumption that the most informative samples are those the classifiers are *least confident* about. The second function is based on the assumption that when the two classifiers have a *high degree of disagreement* about a certain sample, the sample should be labeled by a user.

More formally, let $h_c^1(x_i)$ and $h_c^2(x_i)$ be the two classifiers' confidence scores that sample x_i belongs to the class c based on two different sets of features. The first active sampling function asks for the label of the sample s_j with the lowest prediction score, i.e.:

$$s_j = \underset{x_i}{\operatorname{argmin}}(\max_c h_c^j(x_i)),\ j = 1,2 \tag{5.1}$$

The second active sampling function first finds the conflicts S in the classifiers' predictions:

$$S = \{x_i | \hat{c}_1(x_i) \neq \hat{c}_2(x_i)\} \tag{5.2}$$

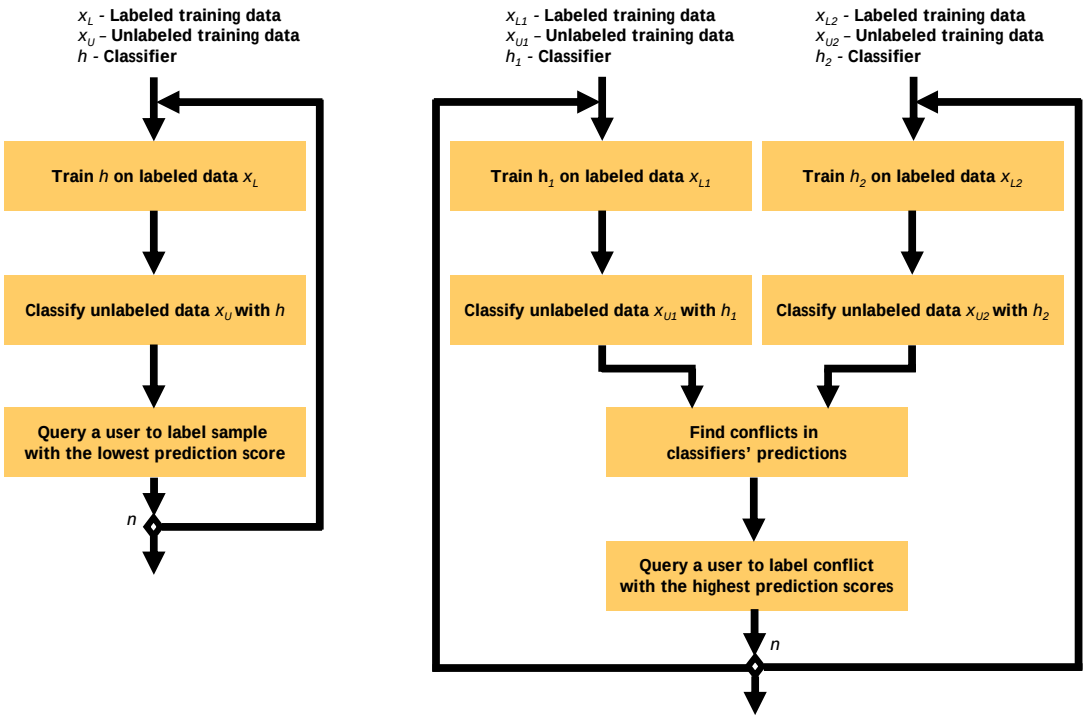
where $\hat{c}_1(x_i)$ and $\hat{c}_2(x_i)$ are predicted classes:

$$\hat{c}_j(x_i) = \underset{c}{\operatorname{argmax}}\, h_c^j(x_i),\ j = 1,2 \tag{5.3}$$

and then chooses for labeling the sample in the set S with the highest confidence score:

$$\underset{x_i \in S}{\operatorname{argmax}}(\max_j h_c^j(x_i)),\ j = 1,2 \tag{5.4}$$

We evaluate the proposed active sampling functions based on the iterative training procedure (Figure 5.3). Again, we use 9-fold leave-one-day-out cross validation and 5 random subsampling rounds. We start with only a few labeled samples, i.e. with 0.3% labeled training data from the previous section. Joint Boosting classifiers on acceleration and infra-red data are then trained and applied to the pool of unlabeled training data. The most informative samples are chosen for labeling by one of the two proposed active sampling functions and added to the labeled training set. The classifiers are then re-trained, and the procedure continues until the size of the labeled training data reaches the size of the four configurations from the previous section, i.e. 1.3%, 2.5%, 6.3%, and 12.5% of 8 days of training data.



(a) Low scores (b) Conflicts

***Figure 5.5**: Active learning algorithms.*

The iterative procedure is shown in Figure 5.5(a) and Figure 5.5(b). In each iteration, the first active sampling function (Equation 5.1) finds two samples for labeling, the one that is predicted with the lowest confidence level based on the acceleration classifier, and the one that has the lowest score based on the infra-red classifier. These two samples are then labeled and added to the labeled training set. The second active sampling function (Equation 5.4) searches the prediction space for conflicts, i.e. samples that are classified differently by classifiers based on acceleration and infra-red data, and chooses for labeling the one that the classifiers predicted with the highest confidence level. That sample is then labeled and added to the set of labeled training data.

5.5.1 Results

Table 5.3 shows the classification results for acceleration and infra-red data, as well as for the classifier combined on these two sensor modalities, after the active sampling labeling process. We compare the results for different amounts of data sampled with the two previously introduced active sampling functions. Additionally, the results are compared with the supervised approach when using the same amount of non-actively (i.e. randomly) sampled labeled training data.

Both active sampling functions outperform the supervised learning approach. On av-

	Acceleration			Infra-red			Combined		
Labeled	Supervised	Active - low scores	Active - conflicts	Supervised	Active - low scores	Active - conflicts	Supervised	Active - low scores	Active - conflicts
1.3%	24.4%	44.5%	47.1%	32.8%	34.5%	29.5%	28.2%	50.9%	51.4%
2.5%	26.9%	52.9%	51.4%	30.5%	39.8%	38.0%	32.3%	59.7%	57.0%
6.3%	35.9%	55.3%	53.8%	30.1%	42.2%	23.7%	39.8%	63.2%	57.5%
12.5%	38.9%	**60.6%**	55.8%	24.5%	**42.3%**	32.2%	35.8%	**64.2%**	63.5%

***Table 5.3**: Comparison of recognition accuracy using 2 different active learning sampling functions and supervised learning for acceleration, infra-red, and combined classifier.*

erage, the first active sampling function for acceleration data based on the low confidence predictions' scores yields 20.6% better accuracy, and the second active sampling function based on conflicts in classifiers' predictions achieves 21.5% better accuracy compared to the supervised case with the same amount of labeled training data. In the case of infra-red data the performance increase is less significant, but still noticeable. Again, we assume that this is due to the noise in the infra-red data introduced by the second subject, which Joint Boosting can not deal with properly. The active sampling function based on the low predictions' scores after labeling 6.3% and 12.5% of training data achieves an accuracy of 42.2% and 42.3%, respectively, which is slightly better compared even to the supervised learning by using 100% of labeled infra-red training data when accuracy is 41.6%.

One must be aware of the potential risk that active learning might focus on the samples that are hard to be learned. It happens occasionally that accuracy decreases by adding more actively sampled labels. For example, when using the active sampling function based on the conflicts for infra-red data accuracy is 38% when 2.5% data is labeled. By continuing the active labeling and reaching 6.3% labeled data, accuracy decreases to 23.7%.

In order to explore the full potential of the multi-sensor approach, in Table 5.3 we also show the performance of the combined classifier, based on the multiplied outputs from the acceleration and infra-red classifiers. That way, we achieve an accuracy of 64.2% when the active sampling function based on the low prediction scores is used and 63.5% when using the active sampling function based on the classifiers' prediction conflicts. In Table 5.3, the best results for acceleration, infra-red and combined classifier are highlighted and the active sampling function based on the low prediction scores consistently performs better, presumably because the active sampling function based on conflicts in classifier's prediction often chooses for labeling the samples close to the decision boundaries.

When comparing the three approaches used in this chapter, one can conclude that the most promising approach is the combined classifier on the actively learned data. Table 5.4 ranks the best results for sensor modalities separately.

Acceleration		Infra-red	
Active - low scores	60.6%	Co-training	45.9%
Active - conflicts	55.8%	Active - low scores	42.3%
Supervised	53.6%	Supervised	41.6%
Co-training	40.7%	Active - conflicts	38.0%
Self-training	40.6%	Self-training	34.4%

***Table 5.4**: Comparison of the best recognition accuracy for all the approaches used.*

5.6 Conclusion

This chapter demonstrated the feasibility of semi-supervised and active learning for reducing the level of supervision in activity recognition.

The two evaluated semi-supervised techniques, self-training and co-training, were found to be capable of learning activity models from a very limited amount of labeled training data. As intuitively assumed, experimental results showed that co-training is less sensitive to mistakes than self-training. It outperforms self-training by augmenting the training process with additional information from complementary sensor modalities. Additionally, in some cases it can achieve higher recognition accuracy than the fully supervised approaches.

The proposed active learning method is based on a pool-based setting where in addition to a small set of labeled training data, there is also a large number of unlabeled training instances available. From the unlabeled pool of data, the algorithm selects the most informative samples to be labeled by a user. We introduced two active sampling functions based on the classifiers' lowest confidence level and on disagreements between the classifiers' predictions. Again, experimental results suggest that it is possible to achieve comparable, or sometimes even higher accuracy than the fully supervised approaches with less labeling efforts.

All this supports the argument that in order to extend the scalability and real-world applicability of the activity recognition systems, one should aim for approaches that require less supervision. In the next chapter, we further explore this promising direction in a more realistic stream-based scenario based on an experience sampling annotation technique.

6

Activity Recognition from Sparsely Labeled Data

Most activity recognition approaches rely on supervised learning methods. However, obtaining substantial amounts of labeled data is often an important bottle-neck for these approaches. In this chapter, we address the labeling challenge by introducing a novel activity recognition approach that requires only coarse-grained annotations. The method utilizes multi-instance learning for activity recognition by representing activity data as bags-of-activities and requiring labels only on the bag level. We also propose several novel extensions of multi-instance learning to support different less intrusive annotation strategies. The validity of the approach is demonstrated on two public datasets for three different labeling scenarios.

6.1 Introduction

As shown in Chapter 2 annotating activity data remains one of the main challenges in real-world settings. The goal of this chapter is to explore an alternative direction for decreasing the level of supervision in activity recognition. In the machine learning community, *multi-instance learning* [Zhou 2004] has emerged as a promising approach for problems with *incomplete* knowledge about labels of training samples. As multi-instance learning is robust to labeling noise and can achieve competitive classification accuracy using only a small amount of training data, we argue that it is excellently suited for the activity recognition problem where labels are hard to obtain and training data is often ambiguously labeled.

Experience sampling (see Chapter 2) enables online annotation of activities during long-term recordings. Thus, we apply our multi-instance approach in this labeling scenario. The two major benefits of the proposed approach are: 1) It enables to decrease the level of annoying experience sampling interruptions by probing a user only occasionally about performed activities and 2) A user does not have to provide accurate start and ending times of activities.

The main contributions of the chapter are threefold. First, we reduce the level of supervision in activity recognition by acquiring knowledge from incomplete labels using multi-instance learning. Second, we extend multi-instance learning to support different annotation strategies by initializing the algorithm with a few correctly labeled data samples and adjusting it to the multi-class setting. Third, we explore and analyze the applicability of the approach in three different labeling scenarios based on an experience sampling annotation method: 1) A user is asked only about his current activity, 2) A user provides information about all activities he performed during a given time interval but without exact time when the activities occurred, and 3) A user provides information only about the activity he was performing most of the time during a given time interval. In all three cases, we achieve high recognition performance with a small amount of experience sampling probes. That way, the level of experience sampling annoyance is reduced. We also report results on two different public datasets in order to avoid bias to a particular dataset only.

The rest of the chapter is structured as follows. Section 6.2 describes the multi-instance learning framework and the applied algorithm. Section 6.3 motivates and presents three different annotation strategies evaluated in this chapter. In Section 6.4 we present the goals of our experiments. Section 6.5, Section 6.6, and Section 6.7 report on our modified extensions of multi-instance learning for activity recognition and experimental results of all three labeling scenarios evaluated in this chapter. Section 6.8 discusses the results of a comparative evaluation procedure. Section 6.9 summarizes our results.

6.2 Multi-Instance Learning

Standard supervised learning requires labeled training data, i.e. each training instance is associated with its corresponding class label. Multi-instance learning relies on a significantly weaker assumption about the labeling information. Here, the labels are not assigned to the individual training instances, but rather to sets of instances, namely *bags-of-instances* (Figure 6.1). The bags are labeled based on the following two rules:

1. The bag is labeled *positive* if *at least one* instance in the bag is positive
2. The bag is labeled *negative* if *all* instances in the bag are negative

Even though the labels of individual patterns are not provided, multi-instance learning successfully copes with the ambiguity of not knowing which of the instances in positive bags are actual positive instances and which ones are not. It aims to find the optimal labeling of the instances in positive bags that consequently leads to the optimal classifier.

6.2.1 Multi-Instance SVM (miSVM)

We have seen in Chapter 2 that discriminative classifiers such as Support Vector Machines (SVM) achieve high recognition performance on activity data. Thus, in our experiments

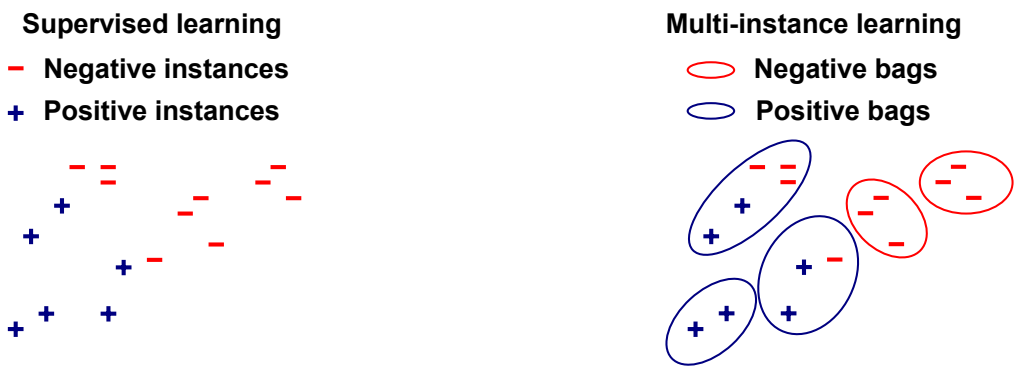


Figure 6.1: *Difference between labeling rules in the supervised and multi-instance settings. For supervised learning labels are provided for each training instance. Multi-instance learning requires labels for sets of instances (so called bags-of-instances).*

we adopt a modified version of the support vector framework that enables activity recognition from sparse labels - called multi-instance SVM (miSVM) [Andrews *et al.* 2003].

In the multi-instance setting the bag labels provide only partial information about the labels of their comprising instances. A negative bag label imposes a negative label for each instance in the bag. In contrast, unique labels of the instances in positive bags are not known. If $y_i \in \{-1,+1\}$ are instance labels and $Y_I \in \{-1,+1\}$ corresponding bag labels, the above mentioned two bag labeling rules can be compactly expressed as:

$$Y_I = \max_{i \in I} y_i \tag{6.1}$$

or alternatively as a set of linear constraints:

$$\sum_{i \in I} \frac{y_i + 1}{2} \geq 1, \ \forall Y_I = 1 \tag{6.2}$$

$$y_i = -1, \ \forall Y_I = -1 \tag{6.3}$$

Intuitively, in order to extend the support vector machine (see Section 3.2.5) setting to the multi-instance case, based on these bag labeling rules, one could construct the maximum-margin hyperplane in the following way: There should be at least one instance from every positive bag in the positive halfspace, while all instances belonging to negative bags are in the negative halfspace (Figure 6.2).

Practically, miSVM treats the labels of instances in positive bags as unobserved indicator variables. The goal is to maximize the regular hyperplane margin (or soft-margin) (Equation 3.12 and Equation 3.13 in Section 3.2.5), jointly over hidden label variables (i.e. possible label assignments) (Equation 6.2 and Equation 6.3) as well as over linear (or kernelized) discriminant functions (i.e. possible hyperplanes). This leads to a hard mixed integer programming problem and therefore, an iterative local optimization heuristic is applied in the following manner (Figure 6.3).

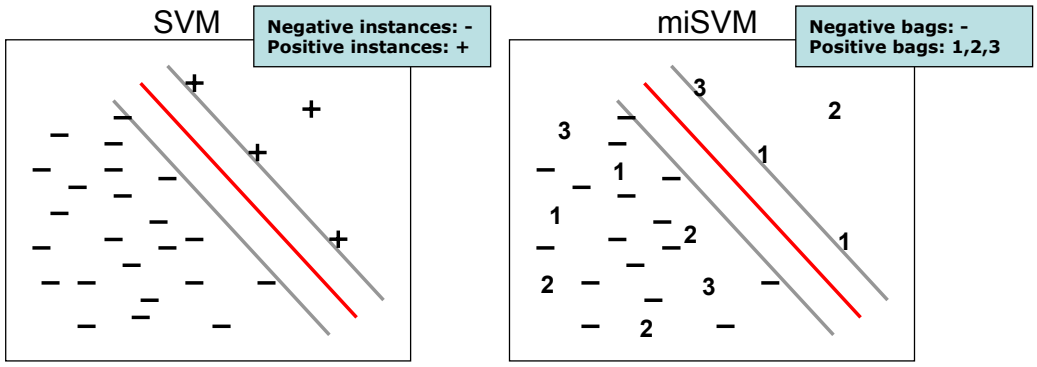


Figure 6.2: *Visualization of the maximum margin SVM classifier extended to the multi-instance setting (miSVM).*

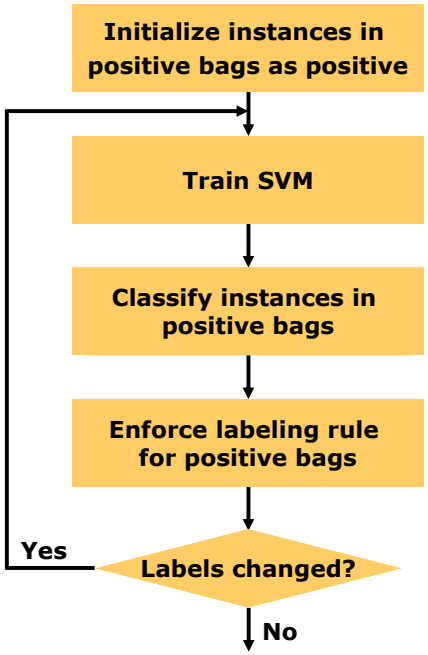


Figure 6.3: *Iterative optimization procedure of miSVM algorithm.*

Initially, all instances in positive bags are assumed to be positive. Given this initial labeling assumption, the regular SVM margin optimization problem is solved. Based on the found maximum-margin hyperplane, instances in positive bags are classified. Furthermore, since positive bags contain at least one positive instance, an additional constraint is applied if the current classifier function classifies all instances in a positive bag as negative. The "least negative instance" is assigned the positive label, i.e. the instance with the highest value of discriminant function. The classifier is then re-trained based on the refined labels and the procedure is iteratively repeated until it converges to a local minimum of the SVM objective function, i.e. until the labeling of instances in positive bags

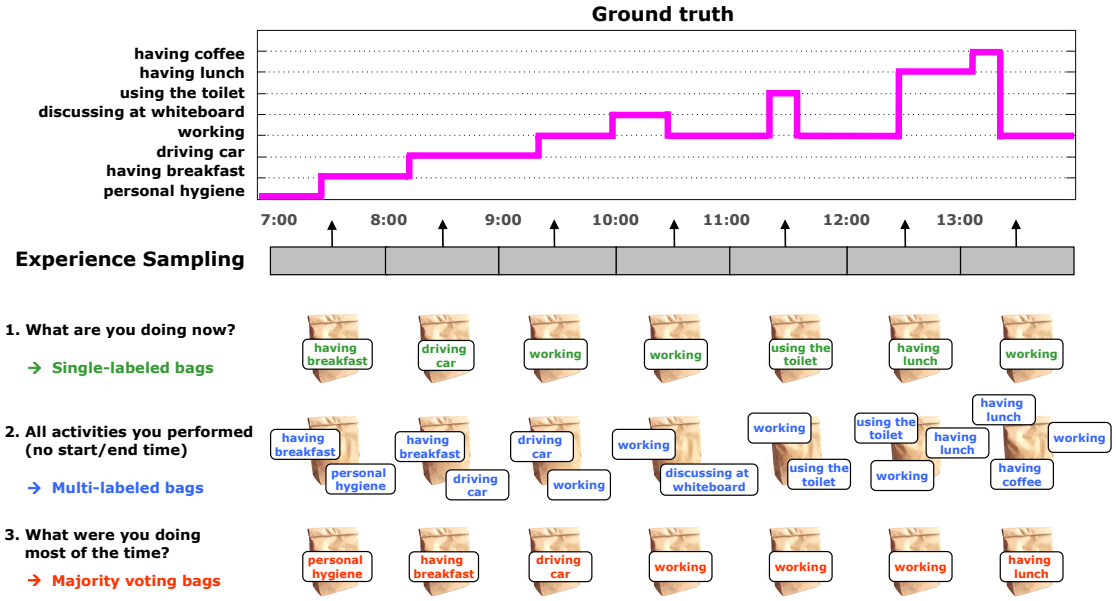


***Figure 6.4**: Illustration of three different bag-of-activities generators: single-labeled bags, multi-labeled bags, and majority voting bags.*

does not change any more. The final classifier can then be used in a standard way.

6.3 Bag-of-activities Generators

In this section, we show how multi-instance learning can be applied to activity recognition by introducing *bags-of-activities*. This way, the activity labels do not have to be provided for each data point but rather on a very coarse level: sensor data is grouped into bags and the labels are provided for the bags. We explore the applicability of multi-instance learning for three different annotation strategies based on experience sampling. It is clearly desirable to decrease the number of experience sampling probes to a minimum. Thus, we aim to prompt a user to provide information about the performed activities as rarely as possible by employing longer experience sampling time intervals. Additionally, the provided level of annotation details is significantly weaker because a user does not have to recall exact start and end times of activities. The bags comprise activity data between the two successive experience sampling prompts. We evaluate the following three bag-of-activities label generators (illustrated in Figure 6.4) that partially fulfill multi-instance bag labeling rules defined in Section 6.2. In order to make multi-instance learning applicable to our three settings, we will discuss our proposed extensions and modifications in Section 6.5, Section 6.6, and Section 6.7.

1) Single-labeled bags. This bag-of-activities label generator is based on user's periodical information about the activity he is currently performing. Thus, in this case each bag

is assigned a single label, i.e. the user's current activity. For longer experience sampling time intervals, bags typically consist of several activities. This introduces noise in bags which can be handled by multi-instance learning. It is highly likely that a user is going to continue the current activity for a certain time after a prompt, i.e. the surrounding instances around the probe have a high probability to belong to the activity provided by a user in the latest prompt. Therefore, we can reduce the level of noise in bags by centering the bags around successive experience sampling prompts and by shortening the bag. This case can be handled directly with a slightly modified multi-instance algorithm. Details will be given in Section 6.5.

2) Multi-labeled bags. In this case, a user provides information about all activities he performed in a given time interval, but not the exact times when the activities occurred. Thus, each bag is assigned multiple labels, comprising all activities the bag is composed of. Unlike the single-labeled bags, where very short events are likely to be missed, the multi-labeled bags allow to recover all activities that took place during the observed time period. The necessary extensions of multi-instance learning for this case will be presented in Section 6.6.

3) Majority voting bags. The third bag-of-activities label generator is based on the assumption that it is easier for a user to recall activities lasting for extended time periods than relatively short activities. Hence, in this case, a user is periodically prompted to provide information which activity he was performing most of the time during a time interval. In other words, the label of a bag is induced by majority voting of individual bag instances. Generating bag labels in this manner creates relatively clean bags, i.e. a large proportion of instances in a bag matches the bag label. The shortcoming of this bag-of-activity label generator is that short activities are often missed, especially when aiming for longer experience sampling time intervals. Further details about the algorithms used for this setting will be described in Section 6.7.

6.4 Evaluation

In this section, we present the goals of our experiments and describe our evaluation procedure.

In our experiments we use two publicly available datasets, namely *PLCouple1* and *TU Darmstadt*, presented in Section 3.1. Since acceleration data in the PLCouple1 dataset is recorded at the relatively high frequency of 20Hz, we extract the following features: *mean*, *variance*, *energy*, *spectral entropy*, *area under curve*, *pairwise correlation between the three axes*, and the first ten *FFT coefficients*, which overall sums up to 144-dimensional feature vectors. It has been shown in [Huỳnh *et al.* 2007] that for high-level activities, it is sufficient to extract features over longer time periods. Thus, we compute

the features over a sliding window of 30 seconds shifted in increments of 15 seconds. In the original paper [Huỳnh *et al.* 2008] where the TU Darmstadt dataset has been introduced, the use of frequency features did not improve recognition performance due to the relatively coarse resolution of acceleration data (2.5Hz). Thus, for this dataset we use, as in [Huỳnh *et al.* 2008] only *mean*, *variance*, and *time-of-day* as features over the same sliding window length used for PLCouple1 dataset producing 13-dimensional feature vectors.

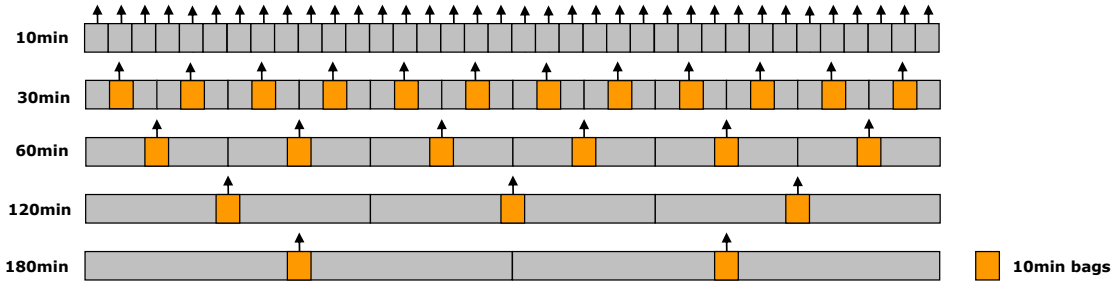


***Figure 6.5**: Different experience sampling time intervals evaluated in the experiments and illustration of 10min bags-of-activities.*

The experiments are designed to systematically compare the three annotation strategies based on experience sampling (described in Section 6.3). We evaluate performance of multi-instance learning for different experience sampling time intervals, starting from 10 minutes to 180 minutes (Figure 6.5). Bags consisting of up to 180 minutes of activity data might include significant amounts of ambiguously labeled data. Besides, it might be hard for a user to recall performed activities during such long stretches of time. Thus, we additionally evaluate performance of the algorithms trained on a shorter 10 minutes time interval of the entire bag centered in the middle of the bag (so called *10min bags* in Figure 6.5). That way, the amount of labeling noise in bags is reduced, but overall amount of training data is also decreased, which makes the recognition more challenging.

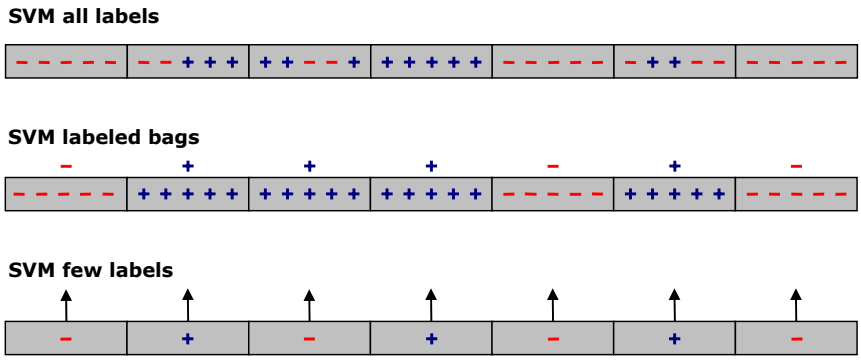


***Figure 6.6**: Three supervised baselines used in the experiments: SVM all labels, SVM labeled bags, and SVM few labels.*

We use several supervised baselines for comparison with the three proposed labeling scenarios (Figure 6.6): 1) the results obtained with standard supervised SVMs using all available training data of individually labeled instances (*SVM all labels*), 2) the results obtained with supervised SVMs using all available training data labeled based on bags' labels, i.e. all instances in negative bags are labeled as negative and all instances in positive bags are labeled as positive (*SVM labeled bags*), and 3) in case of single-labeled bags, the results obtained with supervised SVMs when using only the correctly labeled data points obtained by the single-labeled bags annotation strategy (*SVM few labels*). That way, one can obtain a better understanding of the algorithm's behavior and insights into the benefits and limitations of the approach in the three evaluated settings.

All results reported in the following sections are cross-validated in a leave-one-day-out fashion introduced in Section 3.3.1. In order to extend binary SVMs to our multi-class activity recognition setting, we apply the typical "one vs. rest" approach. In the experiments, we use the Gaussian Radial Basis Function (RBF) kernel (see Section 3.2.5). The kernel parameter γ and misclassification penalty parameter C are determined by coarse grid search over the parameter space. For the fully labeled dataset, we obtained the following parameters: $\gamma = 0.1$, $C = 10$ for the PLCouple1 dataset, and $\gamma = 1$, $C = 10$ for the TU Darmstadt dataset. These parameters are consistently used in all following experiments.

Interestingly, regular SVMs applied to fully annotated activity data obtain slightly better recognition performance compared to the results reported in the original publications where the PLCouple1 and TU Darmstadt datasets were introduced, based on Decision Trees and Naive Bayes classifiers, respectively. With SVMs, we achieve an accuracy of 71.3% on PLCouple1 dataset and 76.3% on the TU Darmstadt dataset. The previously reported results [Logan *et al.* 2007] for the PLCouple1 dataset are 72.2% average ROC area (corresponding to an accuracy of 53.6% reported in [Stikic *et al.* 2008b]). In [Huỳnh *et al.* 2008] an accuracy of 72.7% was achieved on the TU Darmstadt dataset.

In the following sections we introduce our extensions of the multi-instance learning framework and experimentally evaluate their applicability to activity recognition.

6.5 Single-Labeled Bags

In this section, we report on the experiments conducted on the single-labeled bags from Section 6.3. In this case, the label of the bag represents only the activity a user was performing at the moment of the experience sampling prompt. Thus, unlike the standard multi-instance setting, negative single-labeled bags might comprise not only negative instances but also a smaller amount of positive instances. In order to overcome this issue, we make use of the fact that for single-labeled bags, in principle we know to which instance in a bag the provided bag label belongs to. By carrying out the experiments with only that limited amount of accurately provided labeled activity data, standard supervised SVMs already perform surprisingly well. Therefore, we initialize the iterative

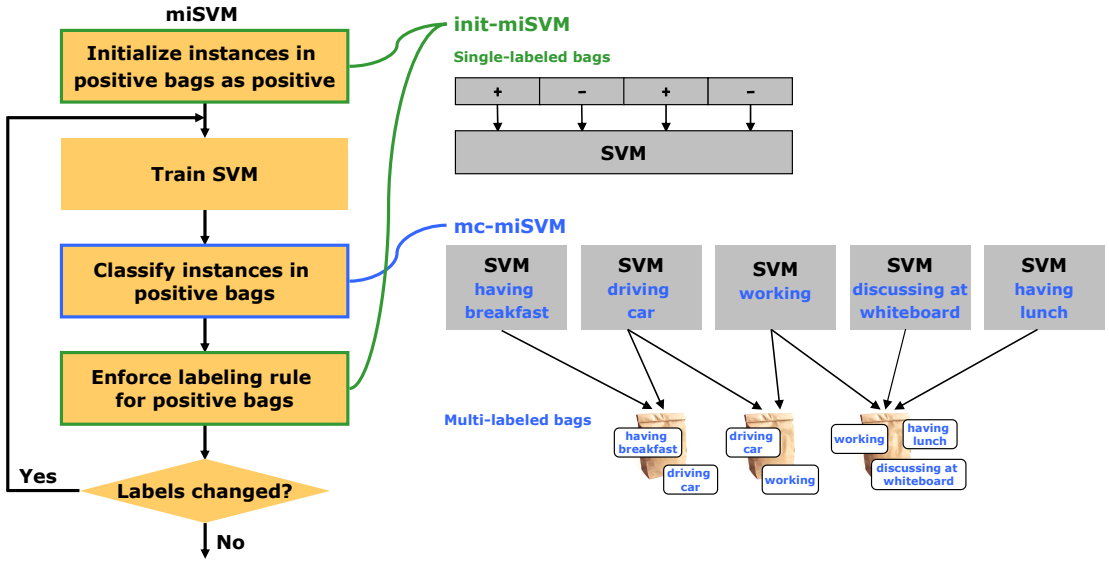


***Figure 6.7**: Illustration of the proposed multi-instance learning extensions for single-labeled bags (init-miSVM) and multi-labeled bags (mc-miSVM).*

multi-instance learning procedure of miSVM presented in Section 6.2 by the SVM model learned in a standard supervised way on that restricted set of provided accurately labeled data. Furthermore, we keep the labels of this data fixed during multi-instance learning. This extended version of the miSVM algorithm is illustrated in Figure 6.7 (*init-miSVM*).

As already mentioned in Section 3.1, the activities in the datasets used in our experiments vary greatly in duration and their frequency of occurrence in real life. In order to compensate for highly unbalanced classes (i.e. activities) in these two datasets, we apply activity priors that could be easily acquired, for example from existing time-use study data [Partridge and Golle 2008]. We utilize the activity priors during both stages of the multi-instance learning miSVM algorithm. For training purposes, we employ different cost factors for misclassification of positive and negative instances based on their prior probabilities. Also, during the classification phase of multi-instance learning, miSVM assigns a constrained proportion of positive and negative labels to the instances in positive bags reflecting the activity priors. In other words, we allow the assignment of positive labels only to the corresponding amount of instances for which SVM provides the highest prediction scores. The rest of the instances in positive bags are classified as negative. That way, we maintain the underlying distribution of activities in our dataset and avoid the risk of large classes dominating the multi-instance learning procedure.

6.5.1 Results

Figure 6.8 illustrates the improvement we can achieve by iteratively assigning labels to the instances in the bags obtained through multi-instance learning. The plots show labeling accuracy of instances in the bags at the beginning and end of our extended multi-instance

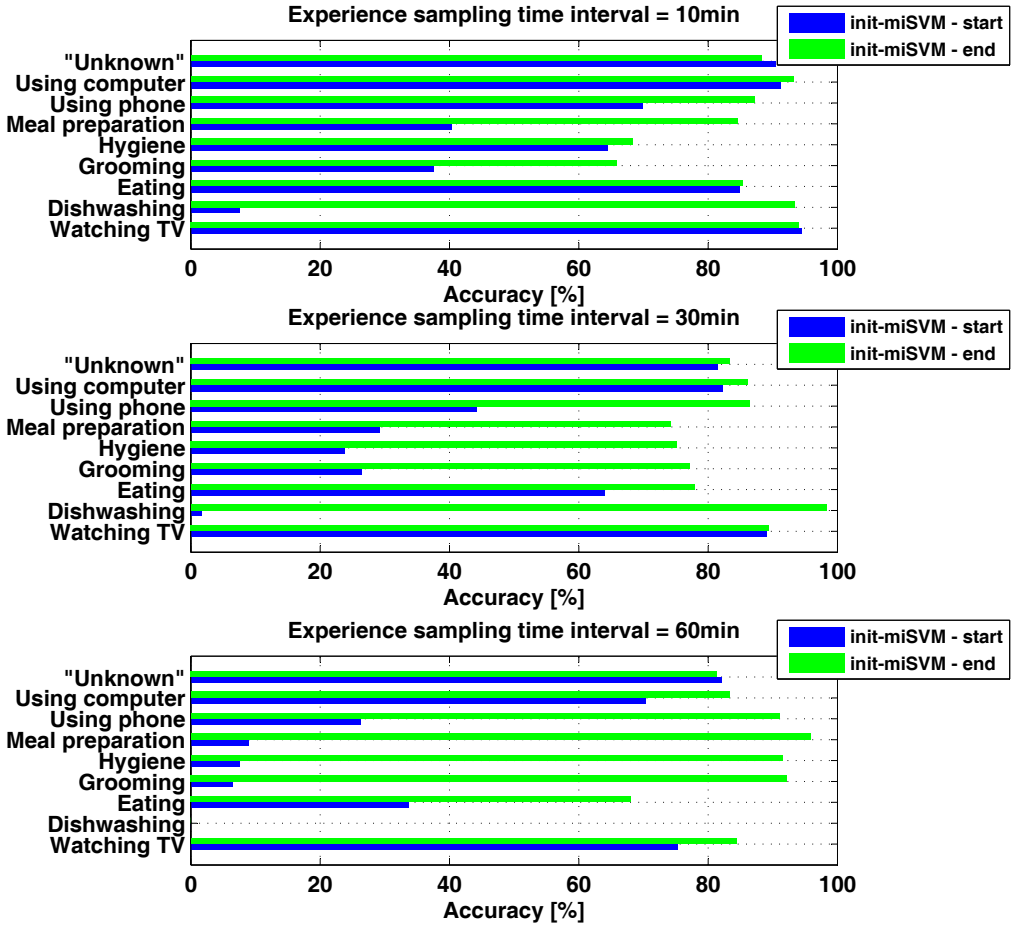


Figure 6.8*: PLCouple1 dataset: Labeling accuracy at the beginning (i.e. first iteration) and end (i.e. last iteration) of multi-instance learning iterative training procedure.*

learning iterative training algorithm for the PLCouple1 dataset for different experience sampling time intervals. From the plots one can observe a consistent improvement for almost all classes of activities. The improvement is impressive for activities *dishwashing*, *grooming*, *hygiene*, and *meal preparation*. When starting with just a few provided labels for these activities, multi-instance learning still enables to find the right labeling of these activities. It can occasionally happen that there are no provided labels for some activities, when using larger experience sampling intervals (e.g. activity *dishwashing* when using a 60min experience sampling time interval).

Figure 6.9 shows the achieved classification accuracy on PLCouple1 and the TU Darmstadt datasets for the standard multi-instance learning miSVM algorithm for both types of single-labeled bags: larger bags (*miSVM labeled bags*) and 10min bags (*miSVM labeled 10min bags*). It also shows the performance of our extended version of the miSVM algorithm, initialized with a few correctly labeled data points, for single-labeled bags (i.e. *init-miSVM labeled bags* and *init-miSVM labeled 10min bags*). The performance of these two algorithms is compared to the following supervised baselines: 1) *SVM*

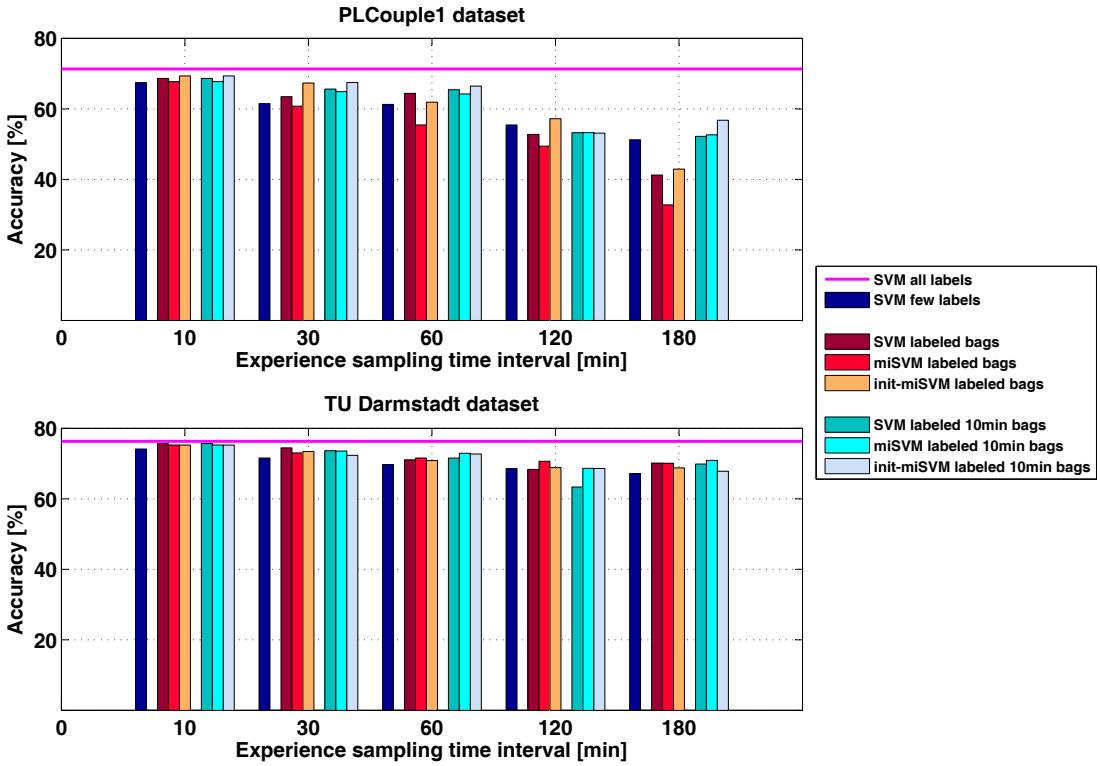


***Figure 6.9**: Single-labeled bags: Comparative performance of supervised baselines (*SVM all labels, SVM few labels, SVM labeled bags, SVM labeled 10min bags*) and multi-instance learning approaches (*miSVM labeled bags, init-miSVM labeled bags, miSVM labeled 10min bags, init-miSVM labeled 10min bags*) for different experience sampling time intervals in case of PLCouple1 (top) and TU Darmstadt datasets (bottom).*

all labels - standard SVM algorithm when using all training data as labeled, 2) *SVM few labels* - SVM algorithm when using only a few correctly labeled data obtained through single-labeled bags annotation strategy, 3) *SVM labeled bags* and *SVM labeled 10min bags* - SVM algorithm when using single-labeled bag labels as labels of all its instances for both types of bags. All algorithms are evaluated for different experience sampling time intervals.

A few trends in Figure 6.9 stand out. First, using experience sampling time intervals up to 60 minutes does not significantly decrease the performance of the algorithms. Thus, the level of interruptions can be significantly reduced by decreasing the number of experience sampling prompts to only one prompt per hour. Second, 10min bags typically achieve competitive (in case of the TU Darmstadt dataset) or even better recognition rates (in case of the PLCouple1 dataset) than the larger bags. These are two very important findings of our experiments that enable an easier way of annotating data for activity recognition.

Furthermore, our modified version of the multi-instance learning algorithm (i.e. *init-miSVM*) improves the performance of the regular *miSVM* algorithm up to 10% for PLCou-

ple1 dataset when using the larger bags. The improvement on 10min bags is not that significant due to the smaller amount of noise in these bags. For the TU Darmstadt dataset, performance is occasionally even decreased due to the poor performance of *SVM few labels* algorithm, which is the initialization part of our *init-miSVM* algorithm. Even though *init-miSVM* typically outperforms *SVM labeled bags* and *SVM labeled 10min bags*, these two supervised approaches still preserve surprisingly high performance, despite of the bag labeling noise.

Lastly, performance of the standard *miSVM* algorithm is often decreased comparing to *SVM labeled bags* baselines. This is most likely due to the fact that the single-labeled bags do not meet the standard multi-instance requirement of clean negative bags. We also conducted an experiment where we discarded the noisy parts of the negative bags, and the performance of the standard *miSVM* algorithm was significantly improved (up to 20% for the PLCouple1 dataset and up to 10% for the TU Darmstadt dataset). In order to further explore this direction in a more realistic setting, we conducted an extensive set of experiments on multi-labeled bags that we report in the next section.

6.6 Multi-Labeled Bags

The multi-labeled bags introduced in Section 6.3 provide complete information about all activities a bag consists of, but without precise assignments of labels to the individual instances. Thus, in this case the multi-instance learning requirement of clean negative bags is fulfilled. As the applied miSVM algorithm utilizes a binary SVM classifier, we are able to deal with multi-labels in a principled way by using the same bag multiple times as a positive bag (i.e. when constructing a binary classifier for each activity that appears in that particular bag).

Since in this case we have information about all activities appearing in the bag, we additionally extend multi-instance learning to the multi-class setting by adjusting the classification phase of the miSVM iterative learning algorithm in the following manner. During classification of instances in a bag, only the classifiers of the activities that are present in that particular bag are allowed to compete for instances in the bag. The classifier that provides the highest prediction score for an instance assigns its label to that instance. That way, we disable unnecessary misclassification of instances by other activity classifiers not included in the bag. This extended version of the miSVM algorithm is illustrated in Figure 6.7 (*mc-miSVM*).

6.6.1 Results

Figure 6.10 shows the results for multi-labeled bags on the PLCouple1 and TU Darmstadt datasets. The plot compares the performance of the following multi-instance learning algorithms: 1) Standard miSVM algorithm on larger multi-labeled bags (*miSVM labeled*

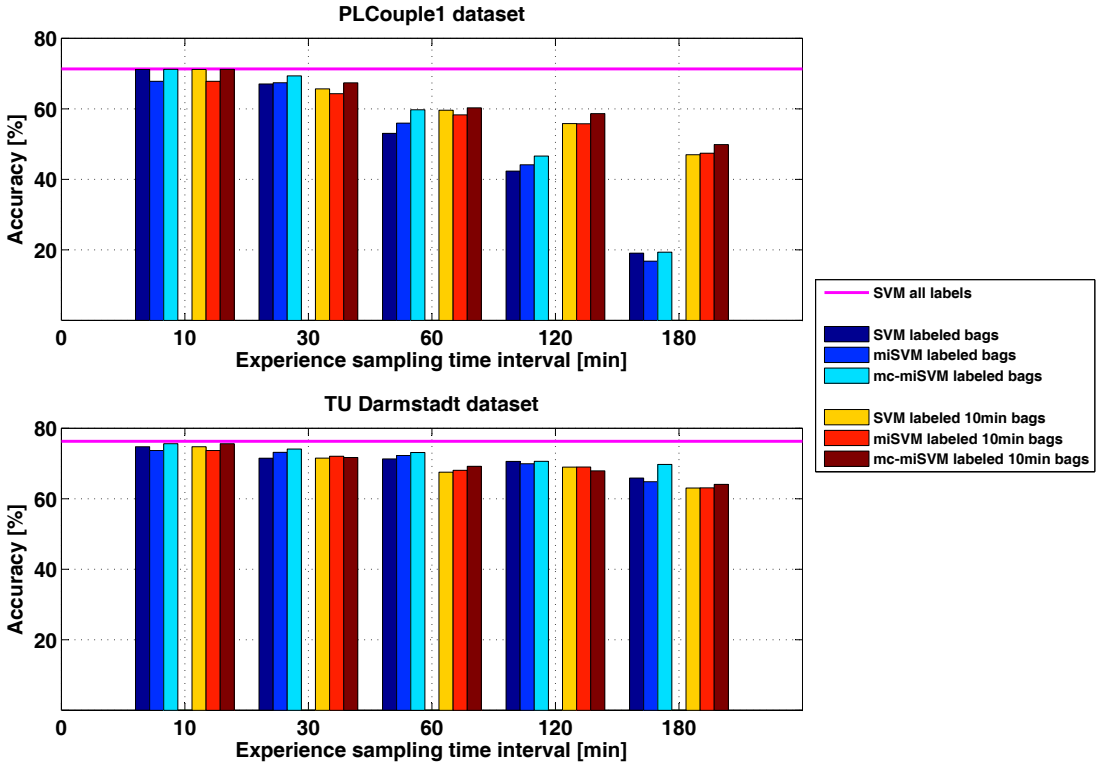


***Figure 6.10**: Multi-labeled bags: Comparative performance of supervised baselines (*SVM all labels, SVM labeled bags, SVM labeled 10min bags*) and multi-instance learning approaches (*miSVM labeled bags, mc-miSVM labeled bags, miSVM labeled 10min bags, mc-miSVM labeled 10min bags*) for different experience sampling time intervals in case of PLCouple1 (top) and TU Darmstadt datasets (bottom).*

bags) and 10min multi-labeled bags (*miSVM labeled 10min bags*), 2) Our extended multi-class version of standard miSVM algorithm both for larger (*mc-miSVM labeled bags*) and 10min bags (*mc-miSVM labeled 10min bags*) to the supervised baselines: 1) *SVM all labels* - standard SVM algorithm on fully labeled data and 2) *SVM labeled bags* and *SVM labeled 10min bags* - SVM algorithm when using multi-labeled bag labels as labels of all its instances for both types of bags.

From the first plot it can be clearly observed that the performance of the multi-instance algorithms does not significantly decrease by using larger experience sampling time intervals compared to the fully supervised baseline (*SVM all labels*). The performance is especially preserved in the case of 10min bags. The only significant drop in the classification accuracy occurred on the PLCouple1 dataset when using the larger bags of 180 minutes experience sampling time interval. This is due to the unbalanced distribution of activities in the dataset. The most frequently occurring activities were included in literally all bags, which consequently led to the complete lack of negative bags for these activities. This effect is less prominent in the 10min bags, which enables competitive recognition rates even when using long experience sampling time intervals. This way, a user is re-

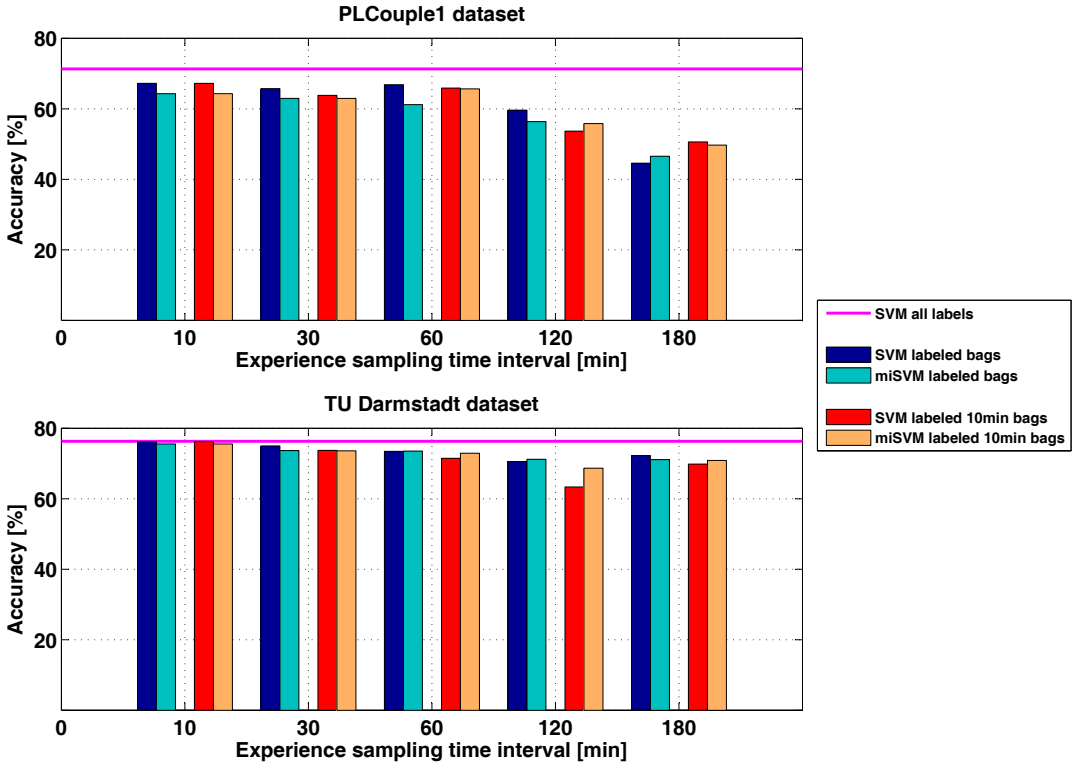


***Figure 6.11**: Majority voting bags: Comparative performance of supervised baselines (*SVM all labels, SVM labeled bags, SVM labeled 10min bags*) and multi-instance learning approaches (*miSVM labeled bags, miSVM labeled 10min bags*) for different experience sampling time intervals in case of PLCouple1 (top) and TU Darmstadt datasets (bottom).*

quired to recall only recent activities occurred in the last 10 minutes. This finding should additionally increase user acceptance of this annotation strategy.

Moreover, the plot shows a clear tendency that our extended multi-class version of multi-instance learning (*mc-miSVM labeled bags* and *mc-miSVM labeled 10min bags*) outperforms the standard multi-instance learning (*miSVM labeled bags* and *miSVM labeled 10min bags*) and the supervised baseline (*SVM labeled bags* and *SVM labeled 10min bags*).

6.7 Majority Voting Bags

Majority voting bags are labeled based on the largest activity appearing in a bag. That can, as in the case of single-labeled bags, introduce a certain amount of noise in negative bags, but of lower significance. The amount of noise in the bags is significantly decreased in this way. Thus, in the following experiments we explore the capability of the standard

version of miSVM algorithm introduced in Section 6.2 to cope with that reduced amount of noise in the bags.

6.7.1 Results

Figure 6.11 shows the performance of the standard multi-instance learning algorithm (*miSVM labeled bags* and *miSVM labeled 10min bags*) and the supervised baselines: 1) *SVM all labels* - standard SVM algorithm on fully labeled data and 2) *SVM labeled bags* and *SVM labeled 10min bags* - SVM algorithm when using majority voting bag labels as labels of all its instances.

As in the previous settings, discussed in Section 6.5 and Section 6.6, again the performance of the algorithms does not decrease when using larger experience sampling intervals up to 60 minutes compared to the fully supervised baseline *SVM all labels*. This clearly indicates the potential for reducing the level of typically annoying experience sampling disruptions.

In this particular labeling scenario, the labels provided on the bag level introduce almost no noise, because the labeling is based on majority voting. Thus, multi-instance learning might not be appropriate for this setting. As can be seen from the plot in Figure 6.11, the performance of *miSVM labeled bags* and *miSVM labeled 10min bags* is often lower than the performance of *SVM labeled bags* and *SVM labeled 10min bags* algorithms. This is due to the fact that multi-instance learning aims at finding the correct labeling of all instances in the bags. But, in this case, large classes dominate the bags, and it is not possible to recover the instance labels of small classes because there are very few, if any, bag labels provided for these small classes.

6.8 Discussion

Table 6.1 and Table 6.2 summarize the best results achieved by the evaluated algorithms in all three explored annotation strategies for the PLCouple1 and TU Darmstadt datasets, respectively.

For single-labeled bags, clearly the best performance is accomplished with *init-miSVM* algorithm, our extended version of the standard miSVM for the PLCouple1 dataset. In the case of the TU Darmstadt dataset, the best results are typically achieved with the *miSVM* algorithm. As already stated in Section 6.5 this is because of the poor performance of the *SVM few labels* algorithm (i.e. initialization phase) on this particular dataset.

For multi-labeled bags, the results indicate that our extended multi-class version of the standard miSVM algorithm, *mc-miSVM*, consistently yields the best recognition rates.

For majority voting bags, surprisingly, the best performance is achieved when applying the regular SVM algorithm on the bag instances without separation of positive and

Time interval	Single-labeled bags		Multi-labeled bags		Majority voting bags	
10min	**69.4%**	init-miSVM	**71.2%**	mc-miSVM	67.3%	SVM
30min	**67.5%**	init-miSVM	**69.4%**	mc-miSVM	65.7%	SVM
60min	**66.5%**	init-miSVM	60.3%	mc-miSVM	**66.8%**	SVM
120min	57.2%	init-miSVM	**58.7%**	mc-miSVM	**59.6%**	SVM
180min	**56.8%**	init-miSVM	49.9%	mc-miSVM	50.6%	SVM

***Table 6.1**: PLCouple1 dataset: Comparison of the best results for single-labeled bags, multi-labeled bags and majority voting bags.*

Time interval	Single-labeled bags		Multi-labeled bags		Majority voting bags	
10min	**75.7%**	SVM	**75.6%**	mc-miSVM	**76.2%**	SVM
30min	**74.5%**	SVM	**74.1%**	mc-miSVM	**75.0%**	SVM
60min	**72.9%**	miSVM	**73.1%**	mc-miSVM	**73.5%**	SVM
120min	**70.6%**	miSVM	**70.6%**	mc-miSVM	**71.2%**	miSVM
180min	**70.9%**	miSVM	69.7%	mc-miSVM	**72.3%**	SVM

***Table 6.2**: TU Darmstadt dataset: Comparison of the best results for single-labeled bags, multi-labeled bags and majority voting bags.*

negative instances in noisy positive bags. Notably, the level of noise in this bag-of-activity label generation strategy is so low that multi-instance learning is not even necessary for achieving high recognition performance. The main drawback of this labeling strategy is that short events are typically missed and the classifier is unable to learn models for these events that might be very important for certain application domains.

We also compare all three annotation strategies (i.e. single-labeled bags, multi-labeled bags, and majority voting bags) by highlighting the recognition results that are in 2% interval around the highest scores (in Table 6.1 and Table 6.2) achieved with different experience sampling time intervals. For both datasets the recognition rates do not vary significantly for different annotation strategies. In the case of the TU Darmstadt dataset, all three annotation techniques perform surprisingly well and by increasing the experience sampling time interval from 10 minutes up to 180 minutes, the recognition performance is decreased by only 4.8%, 5.9%, and 3.5% for single-labeled bags, multi-labeled bags, and majority voting bags, respectively. Thus, our experimental results suggest that competitive recognition performance might be achievable by all three annotation strategies. That can significantly reduce the level of supervision in activity recognition.

In summary, we conclude that it is possible to attain high recognition performance even with very rare experience sampling prompts by extending multi-instance learning to handle different annotation strategies for activity recognition.

6.9 Conclusion

This chapter introduced a novel approach for reducing the required level of supervision in activity recognition. The approach is based on multi-instance learning. It enables automatic recognition of activities from sparsely labeled data.

To this end, we have proposed several novel extensions of multi-instance learning for handling different annotation strategies. We demonstrated the feasibility of the approach on two public datasets for three different labeling scenarios. All three scenarios support less constrained ways of annotating data by requiring neither detailed annotations nor precise start and end times of activities. The experimental results suggest that our extended multi-instance learning algorithms can obtain high recognition performance even though only coarse-grained annotations are provided. We strongly believe that this observation offers a number of still unexplored possibilities for lowering the annotation burden in activity recognition. The results presented in this chapter are but a first step towards better scalability and applicability of activity recognition in real-world settings. In the next chapter, we continue our work in this direction by utilizing a new semi-supervised graph-based approach for activity recognition and evaluating its performance against the multi-instance approach.

7

Multi-Graph Label Propagation for Activity Recognition

This chapter explores another direction for further reducing the level of supervision in long-term activity recognition. We introduce a new activity recognition method that combines small amounts of labeled data with easily obtainable unlabeled data in a semi-supervised learning process. The data are structured in a form of a graph. The labeled data are then used to propagate information through the graph in order to label unlabeled data. We propose two different ways of combining multiple graphs based on data similarity in time and feature space. We evaluate both the quality of the label propagation process itself and the performance of classifiers trained on the propagated labels. Experimental results indicate that the proposed approach outperforms the multi-instance learning approach from Chapter 6.

7.1 Introduction

In this chapter, we show that graph-based semi-supervised learning has the potential to enable recording long-term activity data without the need for detailed continuous activity annotations. It is sufficient to ask users to provide occasional labels about their current activities and use that information together with the remaining unlabeled data for learning activity models. The approach is evaluated in a scenario from Chapter 6 inspired by the experience sampling annotation method. We again aim to decrease the level of experience sampling interruptions by leveraging on the ability of semi-supervised learning to learn from both labeled and unlabeled data. Furthermore, we exploit the fact that in long-term activity recognition the emphasis is on high-level activities that last more than several minutes and can last as long as a few hours. In such setting it is feasible to propagate knowledge through a graph from a few provided labels to the neighboring data that are close not only in feature space but also in time.

The main contributions of the chapter are threefold. 1) We propagate the provided labels to the neighboring data points based on two similarity functions depicted in a graph

structure. The first function is based on the assumption that sensor data of the same activity are similar in feature space. The second function exploits the fact that in long-term activity recognition typical activities span rather long stretches of time allowing to propagate the labels to the neighboring points in the sensor stream. 2) We present two ways of combining the constructed graphs. The first approach incorporates the knowledge from both graphs before the label propagation process takes place. For this a new graph that represents the union of the corresponding graphs is constructed. The second approach combines the knowledge from the graphs after label propagation by confidence voting for labels of the unlabeled data. 3) We evaluate and analyze the quality of the label propagation process itself as well as the performance of classifiers built on the propagated labels.

The chapter is structured as follows. Section 7.2 presents the graph-based approach for reducing the labeling efforts in activity recognition. Section 7.3 discusses the goals of our evaluation procedure. Section 7.4 reports on the experimental results of our algorithm and compares its performance with the approach presented in Chapter 6 based on multi-instance learning. In Section 7.5 we summarize our results.

7.2 Semi-Supervised Label Propagation

In this section, we present our approach of learning from labeled and unlabeled activity data. The approach consists of two steps. In the first step *label propagation* itself is performed, where labels are propagated to the unlabeled data using a graph structure [Zhu and Ghahramani 2002]. As our final goal is to classify unseen activity data, we train, in the second step, a classifier by using both the original labeled data and the new labels obtained during label propagation. This classifier thus incorporates additional information from the training set and should be able to predict activities more accurately than when using only the labeled data.

In order to map the sensor data to a graph, we extract feature vectors from raw sensor data over a sliding window and represent each window by a feature vector. These vectors are the data considered in the graphs.

7.2.1 Label Propagation

First, we will describe the general graph construction procedure and the propagation algorithm. Then we discuss two kinds of similarity measures for graph construction and present two different ways of combining them.

Single Graph Propagation

Intuitively, unlabeled data provides additional information about the underlying data distribution and therefore can be used to guide the label propagation process along high

density regions where the label function is presumably smoother. During the label propagation process, labeled data act like sources that spread their labels through unlabeled data. The basic idea is to build a graph whose nodes are data points (both labeled and unlabeled) and edges encode similarities between nodes. These edges are used for conducting label propagation within the graph.

An important prerequisite for successful label propagation is a graph that properly reflects domain knowledge. Two common ways of constructing graphs are: 1) Fully connected graphs with weighted edges between all pairs of nodes, and 2) Sparse symmetric k-nearest neighbor graphs where nodes i and j are connected by an edge of weight 1 if i is in j's k-nearest neighborhood or vice versa. In our case a k-nearest neighbor graph performed better, presumably due to the removed connections between dissimilar nodes which tend to belong to different classes. Another advantage of sparse graphs is that they are computationally faster than fully connected graphs.

In the following, we present the graph-based label propagation algorithm. If C is the number of classes, let $\{x_1,...,x_l\}$ be the feature vectors representing the labeled data and $Y_L = \{y_1,...,y_l\}, y_i \in \{1,...,C\}$ their corresponding class labels. Let $\{x_{l+1},...,x_{l+u}\}$ be the feature vectors of unlabeled data, with typically $l << u$. The goal is to estimate the labels of the unlabeled data $Y_U = \{y_{l+1},...,y_{l+u}\}, y_i \in \{1,...,C\}$ from all data $X = \{x_1,...,x_{l+u}\}$ and the known labels Y_L.

For that purpose, we assume that an empirical graph $G = (X,W)$ has been constructed. The nodes X consist of both labeled and unlabeled feature vectors. The edges are represented by the weight matrix W, where w_{ij} corresponds to the pairwise similarities of the incident nodes i and j. If the graph is not fully connected, missing edges are associated with weight 0.

Given the weight matrix W, a probabilistic transition matrix T is defined as:

$$T_{ij} = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{ik}}, i,j \in \{1,...,l+u\} \tag{7.1}$$

where T_{ij} can be seen as a probability of passing a label from node i to node j.

Now, let us define a $(l+u) \times C$ label matrix Z. Intuitively, the ith row Z_i represents the probability over the C possible classes for feature vector x_i, i.e. $z_{i,j}$ represents the probability for y_i to be class j.

The propagation is an iterative algorithm which aims at estimating this probability matrix Z. The algorithm initializes the label matrix Z with probability 1 for labels of labeled nodes, i.e. $z_{i,j} = 1$ if $y_i = j$ otherwise $z_{i,j} = 0$, for $i \in \{1,..,l\}$. The remainder of the matrix is initialized to 0. Each label propagation iteration contains three steps. The first step updates the label matrix Z as follows:

$$Z \leftarrow TZ \tag{7.2}$$

In the second step the rows of label matrix Z are normalized to maintain the class probability distributions:

$$z_{ij} = \frac{z_{ij}}{\sum_{k=1}^{C} z_{ik}}, i \in \{1,...,l+u\}, j \in \{1,...,C\} \tag{7.3}$$

The labeled data should be a persistent source of labels. Thus, in the third step we do not allow the initial labels Y_L to fade away by imposing the constraints to the corresponding part of the label matrix Z after the label propagation step (Equation 7.2): for $i \in \{1,..,l\}$ (i.e. labeled nodes only) $z_{i,j} = 1$ if $y_i = j$ and $z_{i,j} = 0$ otherwise. The whole process is then iteratively repeated until convergence i.e. until the label matrix Z does not change anymore.

After the label propagation process, we can assign unlabeled data to their most likely class i.e. the highest class probability score in label matrix Z. More formally, each unlabeled data point x_i, $i \in \{l+1,..,l+u\}$, is assigned label:

$$y_i = \underset{j \in \{1,..,C\}}{\operatorname{argmax}} (z_{i,j}) \tag{7.4}$$

As the activity datasets are typically imbalanced, reflecting different occurrence frequency and duration of activities, it is important to maintain the prior distribution of the activities. Thus, as in Chapter 6, we also include prior knowledge about activity proportions that could be acquired, for example from existing time-use study data [Partridge and Golle 2008]. For that purpose, we incorporate additional constraints in Equation 7.4 by utilizing a multi-class label *bidding* algorithm [Zhu and Ghahramani 2002]. It guarantees that strict label proportions will be met by assigning to each activity class the corresponding proportion of nodes with the highest probability scores for that class.

The algorithm adopts an auction setting, i.e. there is a certain amount of each class labels for sale, and unlabeled data points offer bids for labels based on class probability scores in label matrix Z. More formally, let $P_1,...,P_C$ be the activity priors ($\sum_{i=1}^{C} P_i = 1$). If u is the overall number of unlabeled training data, then uP_c unlabeled training instances should belong to class $c, c \in \{1,...,C\}$. This can be interpreted as uP_c class c labels that we have for sale. Each unlabeled data point x_i, $i \in \{l+1,..,l+u\}$, offers a bid $z_{i,j}$ for class j, $j \in \{1,...,C\}$. The complete set of bids (for all instances and classes) are then sorted and processed from high to low. Let the currently highest bid be $z_{i,j}$. If class j still has labels for sale, a label j is sold to x_i, and x_i quits the bidding. Otherwise, the next highest bid is processed until all unlabeled data points are assigned their estimated classes.

Multi-Graph Label Propagation

A typical similarity function for constructing edges in a graph is *feature similarity*. This similarity function connects two nodes (i.e. feature vectors) in a graph if they are close

to each other in feature space, e.g. based on Euclidean distance. When building such a graph, we connect each feature vector to its k-nearest neighbors by undirected edges expressing symmetric neighborhoods of the incident nodes.

As high-level activities in long-term activity recognition last for extended time periods, we expect data recorded successively to be highly correlated. Thus, time can add valuable information in the propagation process. That is why we also introduce a second type of graph using *time similarity* between nodes. Here, each feature vector is connected to its direct temporal neighbors. Following the linear structure of time, we construct a 2-nearest neighbor graph connecting each node to its predecessor and successor in a sensor stream to represent time continuity of activity data.

These two different similarity measures are complementary and we expect to improve label propagation quality by combining them. For this we propose two different methods:

Union of graphs. The combination takes place before label propagation. We construct a new graph that incorporates knowledge from the respective graphs. This graph is the union [Balcan *et al.* 2005] of feature similarity edges and time similarity edges. The feature similarity graph consists of k nearest neighbors and the time similarity graph contains only 2 nearest neighbors per node. To equalize influence of time and feature edges we weigh the time edges with $\frac{k}{2}$ and the others with 1.

Confidence voting. Here the combination takes place after the propagation process. The label propagation is first performed on each graph separately, and distinct sets of class confidence values are produced. The final score is obtained by summing class confidence scores from both graphs.

7.2.2 Classification

After the label propagation process, we use both the initial labeled training set and the propagated labels to train the classifier. This classifier can be applied to any new data. We use SVM (see Section 3.2.5) in the classification step of our approach. We apply the "one vs. rest" procedure for our multi-class setting and we use a Gaussian radial basis function (RBF) kernel.

As the propagated labels are not perfectly accurate, we take that into account when training the SVM classifier. This is done in two ways. First, at the end of the label propagation process, each unlabeled instance is assigned a confidence score $z_{i,j}$ for its label based on the label matrix Z and the bidding algorithm from Section 7.2.1. We use this information during the SVM training by weighting each instance based on its confidence score. Second, we allow for misclassified training data during SVM training by controlling the trade-off between training error and margin through the SVM misclassification penalty parameter C.

7.3 Experimental Setup

To compare the proposed approach with the multi-instance approach from Chapter 6, we use the same datasets, PLCouple1 and TU Darmstadt, introduced in Section 3.1 in the following experiments. We reproduce the setting from Chapter 6 by using the same set of features extracted over the same sliding window length and cross-validating the results in a leave-one-day-out manner to generate independent test data (for further details see Section 6.4 and Section 3.3).

The aim of the experiments is to reduce the level of experience sampling interruptions. As the used datasets provide fully labeled data, we evaluate our approach for different amounts of labeled data obtained using experience sampling time intervals (as in Chapter 6), starting from 10min up to 180min. We compare the performance of the algorithm to two supervised baselines introduced in Chapter 6: 1) When using all training data as labeled (*SVM all labels*), and 2) When using for training only labeled data provided by experience sampling (*SVM few labels*). We additionally evaluate the *quality of labels* obtained by the label propagation process. The results are also compared to the results of the *single-labeled bags* experiment in Chapter 6. In particular, we compare our approach to the multi-instance learning algorithms: 1) *miSVM* - multi-instance version of the standard SVM classifier, and 2) *init-miSVM* - multi-instance SVM algorithm initialized with labeled data in the first iteration. The other annotation strategies proposed in Chapter 6 are not applicable in this work.

7.4 Results

In this section we report on: 1) the quality of labels obtained by the label propagation process, and 2) the recognition results achieved by a classifier trained on the initially labeled data and the propagated labels. We also compare the results of both stages of the proposed approach to the multi-instance learning approach from Chapter 6.

7.4.1 Quality of Propagated Labels

We use 10 nearest neighbors for constructing the features similarity graph and 2 nearest neighbors for the time based graph. By using relatively few neighbors, we do not allow connections between distant nodes. However, the graph created that way is composed of a certain number of *disjoint* components. It can happen that a component of a graph consists only of unlabeled data and consequently the labels can not be propagated there. As a result, the amounts of data labeled through different similarity graphs are different making the comparison of the quality of propagated labels difficult. Thus, we perform two different evaluation procedures.

1) *Labeling Accuracy.* We report accuracy of propagated labels excluding the parts of training data where the labels are not propagated. In that case accuracy is calculated based on different amounts of data. Still, we believe the reported results reveal the benefits and shortcomings of different label propagation strategies.

2) *Comparison to Multi-Instance Learning.* We want to compare the quality of propagated labels to the labeling outcome of multi-instance learning (Chapter 6). At the end of the iterative multi-instance SVM training procedure, labels are assigned to all training instances in the so-called bags-of-activities. To obtain a fair comparison, we estimate the quality of the graph-based label propagation process in the following manner: we train an SVM classifier on the initial and graph-based propagated labels, and then classify the complete training data including the parts where we were not able to propagate labels.

Labeling Accuracy

Table 7.1 shows the amount of initial labels used for propagation for different experience sampling time intervals. It also shows accuracies of propagated labels for different label propagation strategies for the PLCouple1 and TU Darmstadt datasets. The highest results per dataset (achieved for different time intervals) are highlighted.

		PLCouple1 dataset				TU Darmstadt dataset			
Time interval	Amount of initial labels	Features	Time	Union	Confidence voting	Features	Time	Union	Confidence voting
10min	2.5%	83.7%	**89.1%**	87.1%	88.5%	80.2%	**96.5%**	80.9%	94.8%
30min	0.8%	74.8%	79.9%	77.5%	**81.0%**	79.1%	**93.4%**	79.1%	92.2%
60min	0.4%	65.3%	**71.8%**	69.5%	69.2%	79.7%	**90.7%**	79.2%	90.1%
120min	0.2%	60.0%	**60.3%**	60.1%	59.3%	80.7%	84.3%	80.8%	**86.4%**
180min	0.1%	55.4%	52.4%	**55.5%**	50.1%	81.9%	80.4%	81.7%	**84.1%**

***Table 7.1**: Accuracy of propagated labels for the PLCouple1 dataset and the TU Darmstadt dataset*

For both datasets, the accuracy of time based label propagation is significantly higher than the accuracy achieved by propagating labels based on feature similarity. Only in the case of large experience sampling time intervals of 180min, feature based label propagation performs better. As the average duration of most activities in both datasets is less than 180min, time based label propagation introduces many false labels due to many activities occurring in between two successive experience sampling prompts when labels are provided. Feature based label propagation copes with that better by taking into account feature similarity.

Nevertheless, the achieved accuracy in both cases is impressive considering the amount of labels provided. By 10min experience sampling time interval, 2.5% of training data is initially labeled and we achieve label propagation accuracy of 89.1% and 96.5% for the PLCouple1 and TU Darmstadt datasets, respectively. In case of the 180min time interval,

only 0.1% of training data is initially labeled and the achieved accuracy is 55.5% for the PLCouple1 dataset and 84.1% for the TU Darmstadt dataset. That way, the level of experience sampling interruptions is reduced significantly, and still we are able to accurately propagate labels.

From Table 7.1 one can observe that the accuracy of propagated labels is higher for the TU Darmstadt dataset than for the PLCouple1 dataset. That is due to the fact that some of the activities in the TU Darmstadt dataset (e.g. *driving bike*, *driving car*, *sitting/desk activities*, and *lying/using computer*) last for longer periods and time based label propagation naturally lends itself to these long-term activities. Furthermore, as already mentioned in Section 3.1 after inspection of sensor data in both datasets, we noticed that the PLCouple1 dataset includes a larger number of gaps in the data than the TU Darmstadt dataset. We assume this is due to the fact that the PLCouple1 dataset was recorded only at times when the subject was at home in order to obtain video and audio recordings for offline annotation. The TU Darmstadt dataset was recorded throughout the whole day as the data were annotated by the subject online.

Multi-graph label propagation (i.e. union of graphs and confidence voting in Table 7.1) occasionally outperforms single graphs (i.e. features and time in Table 7.1). Confidence voting typically achieves higher accuracy than union of graphs. As it combines the single graph confidence scores after label propagation it avoids guiding the process in the wrong direction by some misleading features or time based edges in the union of graphs.

One of the benefits of multi-graph label propagation is the following. When using multi-graph label propagation, the number of the isolated components where the labels are not propagated is reduced almost by half. It would also be possible to decrease the number of isolated components by using larger number of neighbors when constructing a graph, but we experimentally observed that accuracy of the propagated labels decreases in that case, since more distant neighbors are used also.

Comparison to Multi-Instance Learning

Figure 7.1 compares the quality of propagated labels for graph label propagation (based on *features*, *time*, *union of graphs*, *confidence voting*) and multi-instance learning (*miSVM*, *init-miSVM* - Section 7.3) for different experience sampling time intervals for the PLCouple1 and TU Darmstadt datasets. One can observe a superiority of graph-based label propagation compared to multi-instance learning. Although graph label propagation based on feature similarity does not necessarily outperform multi-instance learning, the accuracy is significantly improved when combined with time by confidence voting. That latter approach consistently yields the highest accuracy. The improvement is larger for longer time intervals of 120min and 180min (up to 9.5% and 6% for the PLCouple1 and TU Darmstadt datasets, respectively) allowing to further reduce the number of experience sampling interruptions.

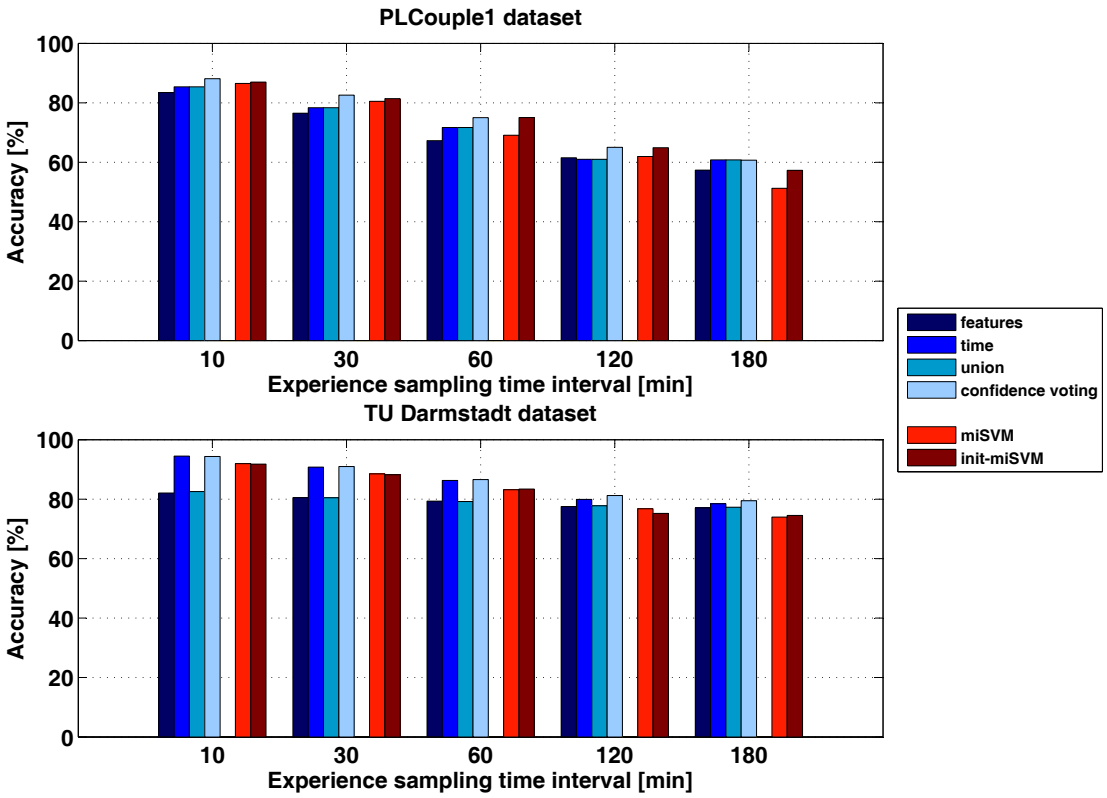


***Figure 7.1**: Comparison of the quality of propagated labels for the graph-based approaches (*features, time, union, *and* confidence voting*) and the multi-instance learning algorithms (*mi-SVM *and* init-miSVM*) for the PLCouple1 and TU Darmstadt datasets.*

7.4.2 Classification Results

In the following, we report on SVM classification results of the left out day's data for graph-based label propagation, and compare its performance to the supervised baselines and multi-instance learning classification results.

Figure 7.2 shows the achieved classification accuracy on the PLCouple1 and TU Darmstadt datasets for the SVM classifier trained on the initial and the labels obtained by different graph-based label propagation strategies: *SVM-features*, *SVM-time*, *SVM-union*, *SVM-confidence voting* for different experience sampling time intervals. The performance is compared to the supervised baselines: 1) *SVM all labels* - SVM classifier trained on fully labeled training data and 2) *SVM few labels* - SVM classifier trained only on a few labeled data points obtained through experience sampling (i.e. the ones we use for graph label propagation). We also compare the performance of our graph-based approach to the results of multi-instance learning: *miSVM* and *init-miSVM* (Section 7.3). A few trends in Figure 7.2 stand out.

First, the graph-based approaches show surprisingly little loss in accuracy comparing to the fully supervised approach *SVM all labels* (Figure 7.2). For larger time intervals of

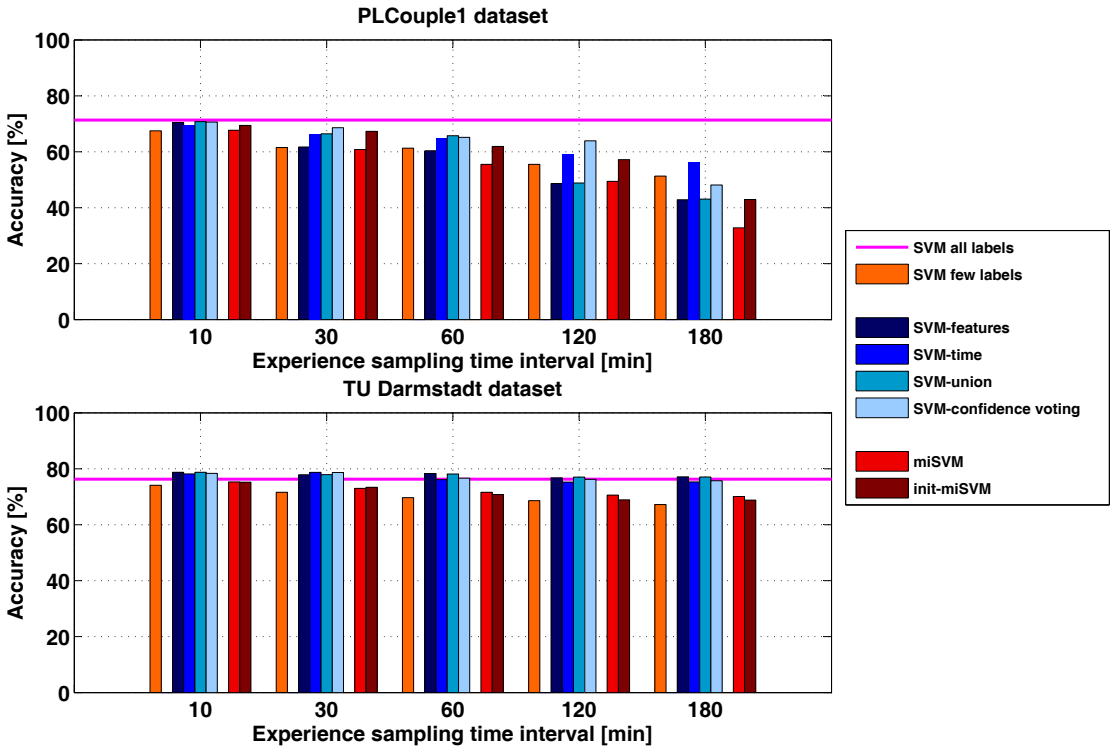


***Figure 7.2**: Comparison of recognition accuracy for the graph-based approaches (*SVM-features, SVM-time, SVM-union, *and* SVM-confidence voting*) to the supervised baselines (*SVM all labels *and* SVM few labels*) and the multi-instance learning algorithms (*miSVM *and* init-miSVM*) for the PLCouple1 and the TU Darmstadt datasets.*

120min and 180min when the accuracy of propagated labels is decreased from 89.1% to 55.5% for the PLCouple1 dataset (Table 7.1) and from 96.5% to 84.1% for the TU Darmstadt dataset (Table 7.1) the training data include a significant amount of noisy labels. We cope with that problem by decreasing the SVM misclassification penalty C from 10 (which is used in Chapter 6 for supervised baselines) down to 0.1 in such cases. That way, we achieve high recognition performance up to 120min for the PLCouple1 dataset when the recognition accuracy is 63.9% by using combined label propagation based on confidence voting. In case of the TU Darmstadt dataset, high performance is retained even for 180min time intervals when the achieved accuracy is 77%. However, in that case we are lacking the labels for some of the very short activities that are completely missed by infrequent experience sampling prompts. Interestingly, the graph based approaches sometimes outperform the fully supervised approach *SVM all labels* on the TU Darmstadt dataset. We will further explore this phenomenon in Section 7.4.3.

Second, we outperform the supervised baseline *SVM few labels* trained only on labels provided by the experience sampling annotation method. This indicates that unlabeled data can be successfully used together with labeled data for training a more expressive classifier. For the TU Darmstadt dataset, our graph methods (*SVM-features*, *SVM-time*, *SVM-union*, and *SVM-confidence voting*) outperform *SVM few labels* on average by 4.4%

for 10min time intervals and up to 9.1% for 180min time intervals. For the PLCouple1 dataset when using larger time intervals of 120min and 180min, *SVM-features* does not outperform *SVM few labels*, but *SVM-time* still outperforms *SVM few labels* by 3.7% and 5.1%, respectively. The combined approaches *SVM-union* and *SVM-confidence voting* are handicapped by the relatively low performance of *SVM-features*. *SVM-confidence voting* handles this successfully and outperforms *SVM few labels* by 8.4% when using 120min time intervals. As a general trend, PLCouple1 exhibits best results for time graphs and confidence voting combination, while the TU Darmstadt dataset shows often the best results with feature graphs, or union of graphs.

Third, compared to multi-instance learning (*miSVM* and *init-miSVM*) the graph-based approaches (*SVM-features*, *SVM-time*, *SVM-union*, and *SVM-confidence voting*) yield better performance for the TU Darmstadt dataset. The average improvement varies between 3.3% for the 10min time intervals over 6.1% for 60min to 7.5% for 180min. This increase in performance for longer time intervals enables to further reduce experience sampling interruptions by using graph-based label propagation. For the PLCouple1 dataset, *SVM-features* is sometimes less effective than *init-miSVM* (e.g. for time intervals of 30, 60, and 120 minutes). However, in combination with the extremely effective time label propagation based on *SVM-confidence voting*, we consistently outperform both multi-instance learning algorithms. The improvement is up to 15.3% for *miSVM* (for 180min time interval). Comparing to *init-miSVM*, the improvement is less significant, but still noticeable (up to 6.7% for 120min time interval).

In Chapter 6, we introduced so-called *10min bags-of-activities* for reducing noise in the bags by considering only data in a shorter 10min time interval around the experience sampling prompt. For complete comparison to multi-instance learning, we carried out the experiments of propagating labels to 10min time intervals around the provided label based on the graph approach. The results are shown in Figure 7.3. The plots compare the performance of the graph approaches (*SVM-features*, *SVM-time*, *SVM-union*, and *SVM-confidence voting*) to the supervised baselines (*SVM all labels* and *SVM few labels*) and multi-instance learning (*miSVM* and *init-miSVM*).

In this case the performance is still very high, but not necessarily higher than in the previous setting due to the following trade-off. We propagate labels to a very short time interval of 10min. After examination of the propagated labels, we observed that their accuracy is very high (up to 89.3% and 98% for the PLCouple1 and TU Darmstadt datasets, respectively). However, the amount of training data is significantly reduced making the training challenging. For the TU Darmstadt dataset and larger time intervals of 120min and 180min the algorithm does not anymore outperform *SVM all labels*. However, *SVM few labels*, *miSVM*, and *init-miSVM* are outperformed, on average by 5.1%, 3.1%, and 4.1%, respectively. For the PLCouple1 dataset, the graph approach consistently yields higher accuracy than *SVM few labels* and *miSVM*. Nonetheless, the recognition accuracy of *init-miSVM* is occasionally slightly higher, e.g. for 30min and 180min time intervals. In comparison to the previous setting, the performance of all graph approaches is more consistent (i.e. there is less variability in the performance) demonstrating the robustness of the approach due to the stronger data similarities both in time and feature space. We

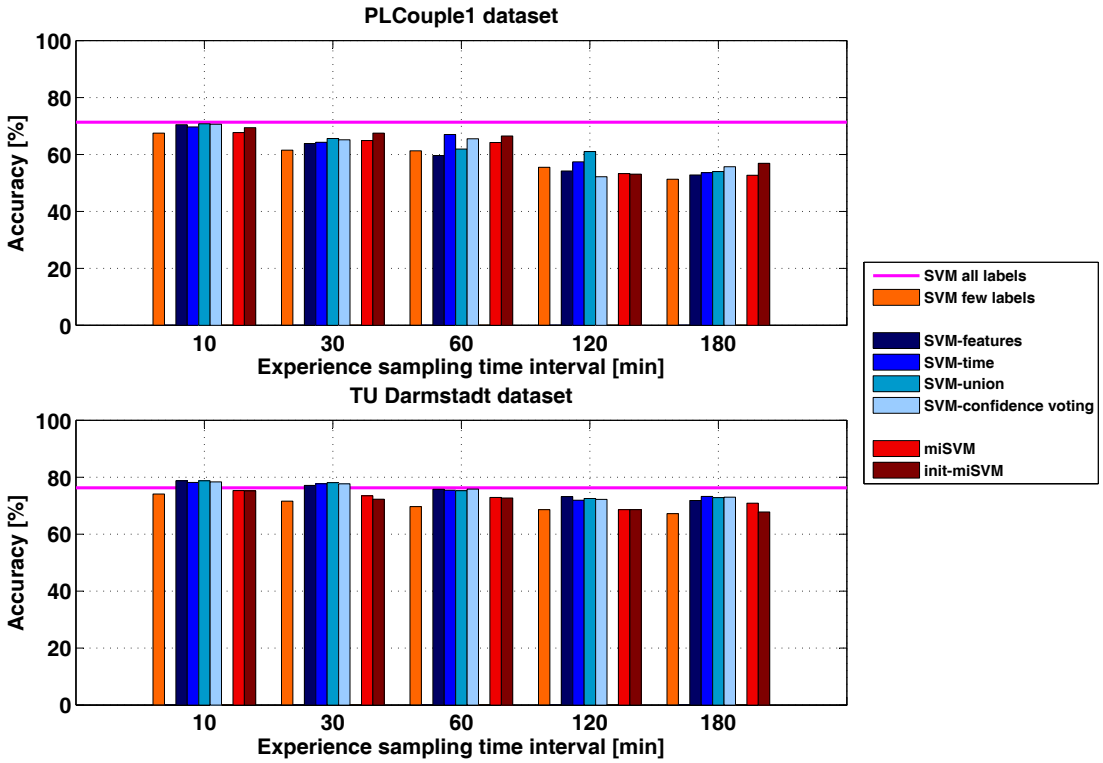


***Figure 7.3**: Comparison of the graph-based approaches to the* "10min bags-of-activities" *multi-instance learning approach.*

expect that in the availability of more training data (i.e. more *"10min time intervals"*) the benefit of this promising approach of label propagation to only small amounts of data around the provided label would be more pronounced.

7.4.3 Discussion

In the following, we further explore and analyze the advantages and limitations of our activity recognition approaches.

One limitation of the graph based approaches is their computational cost. However, in our experiments we utilize k-nearest neighbor graphs. Their advantage is the lower computational cost of the label propagation process than in the case of fully connected graphs. Typically, training of the SVM classifier lasted longer than label propagation. The graphs consist of 9614 and 16875 nodes on average for the PLCouple1 and TU Darmstadt dataset, respectively. In the experiments, the label propagation algorithm converged on average after 71 and 105 iterations for these two datasets. However, in case of very large number of nodes, the algorithms would require significantly larger amounts of memory. That would decrease the efficiency of the algorithm.

So far, we have used overall accuracy as a figure of merit, which is typically utilized as

an overall measure of performance of the multi-class activity recognition systems. However, as the publicly available activity datasets used in this thesis have a very imbalanced character, it is also interesting to analyze average results (i.e. the mean of per class accuracies) and per class accuracies which may allow a better insight into the challenges of real-world activity recognition.

Table 7.2 shows average classification accuracy for both datasets and the different experience sampling time intervals for the compared algorithms, namely: 1) the baselines (*SVM all labels* and *SVM few labels*), 2) the multi-instance learning methods considered in the single-labeled bags scenario (*miSVM* and *init-miSVM*), and 3) the graph based label propagation methods (*features*, *time*, *union*, *confidence voting*). From this table, we can make the following observations.

	Time	SVM	MIL		Graph based approaches			
	Interval	few labels	miSVM	init-miSVM	features	time	union	conf.voting
PLCouple1	10min	25.2%	24.8%	**25.4%**	25.1%	24.7%	25.0%	25.2%
	30min	21.8%	22.9%	**25.1%**	20.9%	23.1%	22.6%	24.1%
	60min	22.8%	20.9%	**23.7%**	20.0%	22.7%	21.8%	22.9%
	120min	19.1%	17.7%	**22.8%**	16.7%	20.2%	15.3%	21.5%
	180min	**17.1%**	12.7%	15.4%	13.1%	17.0%	15.5%	14.4%
TU Darmstadt	10min	15.6%	19.2%	19.8%	13.4%	18.2%	13.4%	19.0%
	30min	12.2%	16.1%	**17.6%**	13.3%	16.6%	13.1%	16.9%
	60min	11.2%	12.4%	12.6%	**16.5%**	14.7%	15.6%	13.2%
	120min	8.3%	**10.5%**	9.6%	9.3%	8.6%	9.1%	9.5%
	180min	8.1%	8.5%	8.2%	9.1%	8.6%	**10.1%**	9.0%

***Table 7.2**: Average classification accuracy for different approaches.* SVM all labels *baseline achieves average accuracy of 25.8% and 20% on the PLCouple1 and TU Darmstadt datasets, respectively. (bold numbers correspond to the best per line)*

Comparing to the previous results (shown in Figure 7.2), we can observe smaller overall values. This implies a large disparity between different activity classes, as the new measure emphasizes the behavior of small classes. The *SVM all labels* baseline is trained on the full set of correct annotations. Thus, we consider it as a theoretical upper bound. Surprisingly, even this fully supervised baseline obtains average accuracy of 25.8% for the PLCouple1 dataset (which consists of 9 classes) and 20% for the TU Darmstadt dataset (which consists of 21 classes).

Still, most of the observations obtained from overall accuracy results hold. For the PLCouple1 dataset, time based graphs and the confidence voting combination obtain better results than features and union of graphs. For the TU Darmstadt dataset, we again observe the small superiority of features and union of graphs. However, in this evaluation framework, the multi-instance learning methods exhibit comparable or sometimes even better results than the corresponding graph methods, in particular for the PLCouple1 dataset. It appears that multi-instance learning deals better with the small classes, as we will show below.

A final difference is that for the TU Darmstadt, the fully supervised baseline *SVM all labels* (caption of Table 7.2) outperforms all other methods. This is an expected behavior, but it was not observed for overall accuracy results. Interestingly, in that case, the

graph approaches were occasionally outperforming *SVM all labels* presumably because the graph methods concentrate more on larger classes during label propagation. We aimed to overcome this problem by introducing activity priors. However, due to the huge class imbalance, this challenge remains an open issue.

In Figure 7.4 and Figure 7.5 we show an example of the propagated labels for the PLCouple1 dataset based on features and time for experience sampling time intervals of 10min and 60min, respectively. The plots also show the ground truth for this cross-validation round. From the plots one can observe that both label propagation based on features and time does not work perfectly well. Feature based label propagation often has false positives and time based label propagation typically does not work very well at the transition between the activities. One way to overcome these problems is our multi-graph label propagation based on the union of the graphs and confidence voting.

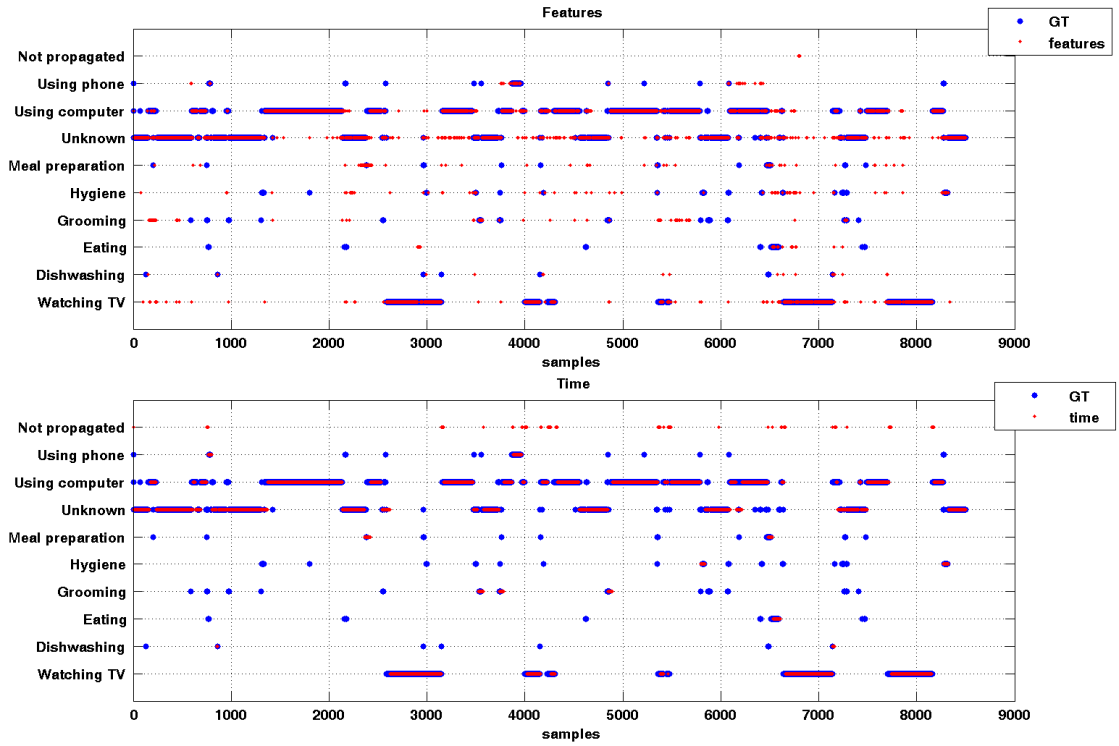


Figure 7.4: *The plots show ground truth (GT) and propagated labels for feature and time label propagation on the PLCouple1 dataset for experience sampling time interval of 10min.*

The quality of propagated labels in Figure 7.4 and Figure 7.5 for larger activity classes such as *using computer* or *watching TV* is better than for smaller activity classes such as *eating* or *hygiene*. For larger experience sampling time intervals of 60min, very short activities (e.g. *dishwashing*) are completely missed by experience sampling prompts. Furthermore, for larger experience sampling time intervals of 60min there are more isolated components of the graphs where labels are not propagated than in the case of 10min time intervals.

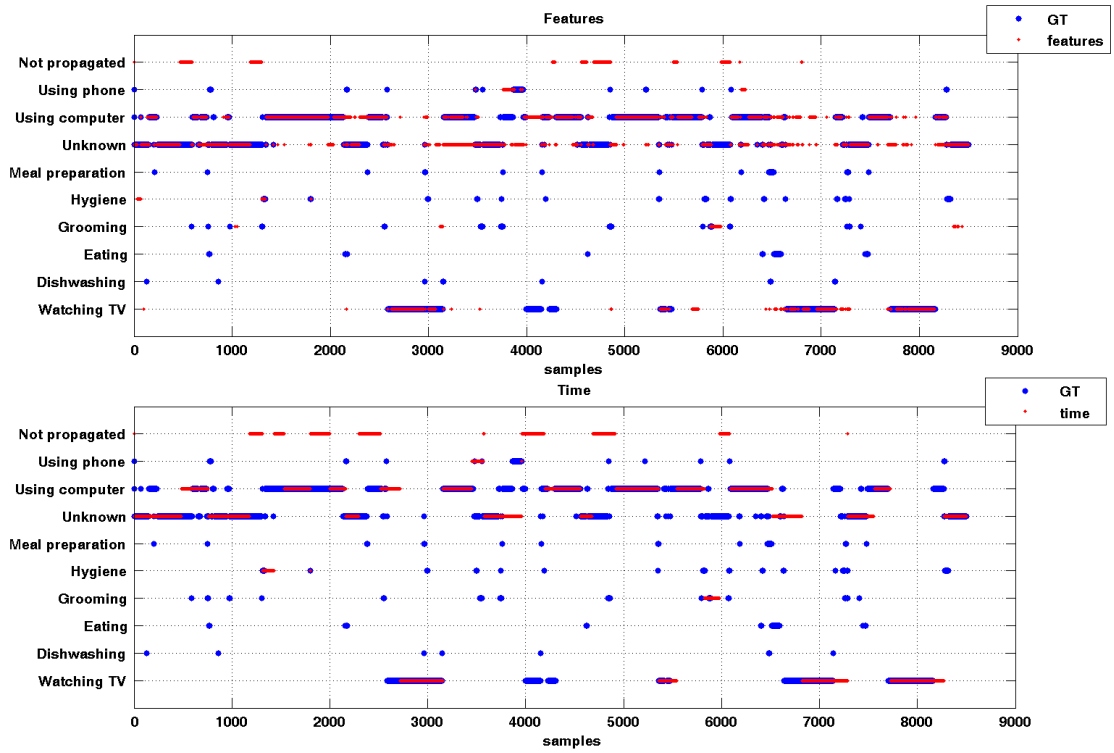


***Figure 7.5**: The plots show ground truth (GT) and propagated labels for feature and time label propagation on the PLCouple1 dataset for experience sampling time interval of 60min.*

In order to further explore the behavior of the different classes, we now discuss the per-class classification accuracy obtained for some of the activities in both datasets for experience sampling time interval of 60min. The amount of data per activity varies between 38.09% and 0.4% for the PLCouple1 dataset and between 61.88% and 0.04% for the TU Darmstadt dataset. Due to such disproportion of classes, even the fully supervised baseline *SVM all labels* obtains largely varying results across activities. Very short activities such as *using phone*, *hygiene*, and *meal preparation* in the PLCouple1 dataset or *discussing at whiteboard*, *walking while carrying something*, and *picking up cafeteria food* in the TU Darmstadt dataset are never recognized. Interestingly, with the proposed methods (i.e. multi-instance learning and graph based label propagation) some of these activities are at least occasionally recognized (e.g. *meal preparation* with miSVM, init-miSVM, and confidence voting in the PLCouple1 dataset or *walking while carrying something* with init-miSVM in the TU Darmstadt dataset). In general, multi-instance learning deals better with small classes than the graph based methods. Furthermore, the results for both the fully supervised and semi-supervised approaches could be potentially improved by including the activity priors in the classification process itself (i.e. during testing phase). Some of the very short activities in the TU Darmstadt dataset (i.e. *driving bike*, *sitting/having a coffee*, *personal hygiene*, and *brushing teeth*) are recognized surprisingly well (accuracy is up to 93.5% for *driving bike*), considering the amount of data

available for training (less than 1% per class). This is presumably due to the regularity of these activities (they occur daily, at relatively regular time), and the *time-of-day* feature helps the classifier to distinguish between classes sharing a common motion patterns, even when only a few labels per class are provided. Furthermore, after investigating confusion matrices, we observed for the TU Darmstadt dataset, many confusions between some of the similar activities, such as *walking*, *walking freely*, or *walking while carrying something*. This suggest that in long-term realistic recordings one should rather aim for definition of activities at the higher level of abstraction, i.e one class for walking might be more appropriate then fine-grained definition of activities. Performance of classifiers might be improved that way, even in the semi-supervised settings. Moreover, annotating data that way might introduce less noise in the provided labels under real-world conditions. For large activity classes such as *using computer* (which represent 38.09% of data) in the PLCouple1 dataset or *sitting/desk activities* (which represent 61.88% of data) in the TU Darmstadt dataset there is almost no loss in accuracy with semi-supervised techniques comparing to the fully supervised baseline *SVM all labels*. In some cases, the results are even increased with these methods. Lastly, for larger experience sampling time intervals, very short activities (e.g. *dishwashing* in the PLCouple1 dataset or *washing hands* in the TU Darmstadt dataset) are completely missed by experience sampling. These activities have no initial labels in the training set, which prevents learning their models.

7.5 Conclusion

This chapter demonstrated the effectiveness of graph-based semi-supervised learning for activity recognition. The method facilitates long-term activity recordings by using experience sampling without detailed annotations by propagating provided labels to the neighboring data. It is sufficient that users sporadically provide information about their current activities.

We proposed and systematically analyzed two different similarity functions for label propagation based on data similarity in time and feature space. The time based label propagation works extremely well and we incorporate it in a multi-graph label propagation scheme to improve the label propagation process. We conducted experiments on two public datasets and compared the performance of the proposed method with multi-instance learning from Chapter 6 that also aims to reduce the level of supervision in activity recognition. The experimental results suggest that the multi-graph based approach consistently outperforms multi-instance learning both in the quality of propagated labels and the classification of test data, especially for smaller amounts of labels. That enables to further reduce experience sampling interruptions by using longer time intervals up to 120min. In comparison to supervised learning, our approach outperforms the supervised approach that uses the same labeled data for training, proving the hypothesis that unlabeled data can also be used for learning more expressive classifiers.

8

Conclusion and Outlook

In this chapter we summarize the main findings and conclusions of this thesis, and outline the possible starting points for future research in the field of activity recognition.

8.1 Summary of Contributions

This thesis investigated several research directions to bridge the gap between the state-of-the-art of activity recognition approaches and real-world deployment of activity recognition systems. Significant progress has been made during the thesis to make automatic recognition of daily activities scalable to large amounts of activities and users. The main findings and achievements of the thesis are as follows.

Multi-sensor approaches are a promising direction for reducing the overall required number of sensors and the level of supervision in activity recognition. Typical daily activities such as personal hygiene, meal preparation, or housekeeping exhibit a large variability across individuals. Thus, multi-sensor approaches lend themselves well to characterization of different activity properties. In this thesis we explored several ways of combining data from three sensor modalities, namely accelerometers, RFID tags and readers, and infra-red motion detectors to infer important activity characteristics such as body-motion, interactions with objects, and the location where the activity is being performed. In Chapter 4 we showed that sensor fusion enables reducing the number of sensors to just a few RFID tags deployed on the key objects and a single wrist-mounted device comprising an accelerometer and an RFID reader. We demonstrated that the integrated approach compensates for the shortcomings of both sensor modalities while significantly improving the activity recognition results. Chapter 5 presented a systematic comparison of four different algorithms based on semi-supervised learning and active learning for reducing the level of supervision in activity recognition. The experimental results indicate that the multi-sensor approaches are more robust to labeling errors by augmenting the learning process with a complementary source of information. Prior to our study, there existed little work on multi-sensor techniques for minimizing annotation overhead in the field of activity recognition.

Reliable activity recognition can be achieved by using only sparsely labeled data. User-annotated activity data are often coarsely and ambiguously labeled. In Chapter 6 we have introduced a novel activity recognition approach that can cope with these issues by employing multi-instance learning. The proposed approach does not require detailed activity annotations and is robust to labeling noise, which leads to more comfortable ways of annotating data. Furthermore, we extended the approach to allow for new labeling scenarios in which the user is probed only occasionally to recall the recent activities he has performed without the need for providing the exact time when the activities have started and ended. We have evaluated the approach on two large public datasets consisting of multiple days of non-staged recordings of up to 20 daily activities. The experiments analyzed the trade-off between labeling efforts and recognition performance. We showed that the extended approach can achieve high recognition rates with very few labels provided (every 60 minutes).

Semi-supervised learning facilitates decreasing the labeling efforts in long-term activity recordings. As many activities of interest are performed on a daily basis, it is relatively easy to obtain large amounts of unlabeled activity data. In Chapter 7 we have investigated to which extent unlabeled data can enhance the training process and whether we can build more expressive classifiers that way. We have introduced a novel graph-based semi-supervised activity recognition approach that propagates information from a few provided labels to the neighboring data points. In order to improve the label propagation process, we constructed multiple graphs based on data similarity in time and feature space for combined label propagation. This way we are able to avoid in part the propagation of labels through some of the misleading graph edges and improve the accuracy of the label propagation process. The comparative evaluation of this approach and the multi-instance approach showed the potential of the graph-based approach to further decrease labeling efforts in long-term recordings. High recognition scores are achieved even when only coarse labels are provided every 120 minutes. All this confirms the hypothesis that semi-supervised learning is an adequate and promising approach for minimizing the labeling burden in activity recognition.

8.2 Conclusion

The work in this thesis contributed to the motivating scenario of automatic health assessment in several ways.

First, the proposed methods enable the robust characterization and recognition of a broad range of daily activities such as ADLs and IADLs which are often used in the medical field for evaluation of an individual's physical and mental skills. The automatic health assessment could in the future better reflect for example whether an elderly person can function well and independently in a daily life. Long-term activity monitoring is believed to be a viable approach for detecting the first changes in behavior indicating

the early signs of different age-related diseases. Therefore, automated detection of daily activities is an important step towards this aim.

Second, the decreased number of sensors necessary for accurate activity recognition improves the wearability and deployment of activity recognition systems. Clearly, sensors should take a minimally invasive and socially acceptable form factor. They should put no constraints on the physical movement and comfort of the user wearing them or on the environment where they are deployed. It is believed that unobtrusive activity monitoring would increase user acceptance of such systems. Thus, reducing the number of the required sensors is a first step towards this goal.

Third, the new more practical methods of labeling activity data presented in this thesis allow for long-term activity recordings without posing a significant labeling burden on a user. Annotation of activity data is one of the major problems in realistic long-term activity recordings. The proposed algorithms enable achieving high recognition scores with less labeling efforts. Furthermore, one of the main findings of our work is that unlabeled activity data carries a wealth of information. That is a useful starting point for further exploration of adopting both labeled and unlabeled activity data in the training process. It should be straightforward to obtain large quantities of unlabeled data recorded by multiple users in real-world settings and combine them with small amounts of labeled data to accommodate for more powerful classifiers and better discrimination between different activities. Unlabeled data could also be used together with the algorithms proposed in the thesis for recognition of different types of activities relevant for other application domains from Chapter 2.

However, additional steps are still necessary to overcome the limitations of the proposed solutions. In the following section, we briefly outline several research directions for potential further improvements.

8.3 Outlook

There are several possible extensions and continuations of the work presented in this thesis.

Further algorithmic improvements. In the following we suggest a few natural extensions of our algorithms. In Chapter 5, we demonstrated the feasibility of multi-sensor approaches based on co-training and active learning to reduce the level of supervision in activity recognition. We believe that a hybrid approach could further enhance the potential of unlabeled training data. In the initial phase the system would actively ask for labels of the most profitable activity samples. In the second phase, co-training could highly benefit from actively learned labels. In Chapter 6 we proposed so-called bags-of-activities and multi-instance learning for activity recognition and introduced several extensions suitable for new labeling scenarios. Further investigation of other multi-instance learning

algorithms based on deterministic annealing [Gehler and Chapelle 2007] could enable soft-assignments of labels to the instances in the bags and more robust iterative labeling process. Moreover, in order to additionally ease the necessary labeling efforts, varying bag sizes could be used where the bag boundaries would be placed at the recognized activity transitions. Similarly, the graph constructed that way would also incorporate the additional segmentation information about the activity transitions. Context-sensitive experience sampling [Intille *et al.* 2003] could be utilized for this purpose. Multi-graph label propagation presented in Chapter 7 offers a number of possible derivations such as for example multi-graphs based on intersection of the corresponding graph edges.

Other interesting approaches to scalable recognition of daily activities and open challenges in real-world settings that may provide significant advances are:

- **Activity spotting.** Activity discovery [Minnen *et al.* 2006a] identifies and models occurrences of short-term motion primitives. As typical daily activities are composed of many sub-activities exhibiting high intra-class variability, a new promising approach for their detection is spotting [Blanke and Schiele 2009] of the most discriminative sub-activities for a particular high-level daily routine. The approach is appealing due to the smaller amounts of data needed and the reduced computational requirements. Thus, it could be beneficial in scenarios where modeling and complete recognition of the entire high-level activity is not necessary but just detection of its occurrence is sufficient.

- **Algorithms for imbalanced datasets.** One of the real-world activity dataset challenges is that they are typically very imbalanced, i.e. the amount of data varies significantly for different activities due to the different average durations and occurrences of activities in real life. This makes the classifier's training more difficult, and up to now this is still an open research question. However, in the machine learning community several ways for dealing with class-imbalance [Liu *et al.* 2009b] have been introduced, such as under-sampling, over-sampling, or cost-sensitive learning. These methods have potential of yielding improvements in recognition performance and it might be worth exploiting their applicability in the field of activity recognition.

- **Algorithms for handling overlapping and interleaving activities.** As stated in Chapter 1, daily activities often happen in an overlapping or interleaving manner and the activity recognition algorithms should be able to cope with that challenge. The joint boosting algorithm in Chapter 5 has been able to successfully deal with multi-labeled data. This issue has been started to be explored in e.g. [Hu *et al.* 2008, Modayil *et al.* 2008]. Another promising approach along these lines has been introduced in [Huỳnh *et al.* 2008]. It is based on topic models and could be potentially able to express overlapping activities as probabilistic activations of different topics.

- **Online adaptive learning.** It would be most desirable to have an adaptable activity recognition system that is able to adjust itself to new users or activities. However,

the current methods require explicit adaptation by changing the training set or the recognition algorithm. Online algorithms, such as online boosting [Grabner and Bischof 2006] could be beneficial for activity recognition allowing for training the classifier online and incrementally as new data becomes available.

Sensors and hardware setup. Hardware design is out of the scope of this thesis. However, during the thesis we have identified numerous limitations of the current sensing platforms. For example, the very short range of the RFID reader antenna caused a significant number of user-object interactions not being detected by a reader and a more appropriate RFID reader antenna range would be highly beneficial for more robust activity inference. Accelerometers are the most common used sensor modality for on-body sensing of user activities. However, other types of inertial sensors such as gyroscopes have proven to provide valuable information for recognition of short-term activities [Zinnen *et al.* 2009] and it is an interesting research question if they are feasible for recognition of typical daily activities. Furthermore, long-term activity recordings in the real-world put additional demands on the robustness and power-efficiency of the deployed hardware. In such settings, sensor failure is clearly more costly.

Additional contextual cues. An important aspect of practical context-aware applications is the fusion of multiple cues of information. Therefore, more elaborate approaches are required in order to include further aspects of user context. In long-term activity recognition location is another important cue that should be exploited in the future. In Chapter 5 we showed that the usage of a coarse indoor location on room level can be used as a complementary information for semi-supervised co-training of the classifiers. It would be beneficial to extend this experiment to outdoor location as well. An interesting study of combining acceleration and GPS data has already been conducted in [Reddy *et al.* 2008] in the context of determining transportation mode of a user. Furthermore, in long-term activity monitoring time becomes an important feature. In this thesis we followed the approach of [Huỳnh *et al.* 2008] by incorporating time-of-day as a component of the feature vector in the experiments with the TU Darmstadt dataset in Chapter 6 and Chapter 7. In Chapter 7 we also made use of time for the label propagation process. That approach worked extremely well and further investigation of this still unexplored research direction by incorporating prior knowledge about typical activity durations is needed. This and many other pieces of useful information could be retrieved from existing time-use studies [Partridge and Golle 2008].

Long-term studies. Future activity recognition systems should be validated on a larger scale and on real-world datasets consisting of long-term activities performed in non-laboratory environments and by the end-users of the system. This is crucial for at least two reasons. First, only then will we be able to get a better insight to which extent the performance of different activity recognition approaches can be generalized under non-constrained settings. Second, potential issues related to user acceptance of the activity

recognition systems might become visible only during such long-term studies. That might impose new requirements and challenges on the current approaches opening potentially new research directions in this field. Thus, we strongly believe that long-term studies are the next most important step towards applicability of activity recognition in real-world scenarios. As such long-term studies typically require a lot of organizational and financial resources, we strongly encourage the researchers working on activity recognition to follow the example of a few research groups which share their datasets with the rest of the activity recognition community. That would also allow for easier comparison of different approaches.

Automatic ADL/IADL assessment. The ultimate goal of ADL/IADL recognition is not only to recognize the activities but also to perform the complete ADL/IADL assessment automatically and detecting even the potentially dangerous changes in a person's behavior. In order to do that, the development of new algorithms seems to be inevitable. Such algorithms should be able to automatically score whether a person can perform the activities independently. Recognition of cooperative multi-person activities could be a first step towards this goal. Having in mind the ADL/IADL scales in Appendix A, detailed recognition of all steps necessary to be executed in order to perform an activity to the end and possible deviations from these steps [Hoey *et al.* 2007] would also support the automatic assessment. The automatic ADL/IADL assessment would enable care-givers to provide better services to the elderly people by reducing the manual assessment overhead. However, this is still an open and with a few exceptions (e.g. [Wilson 2005]), relatively unexplored research question.

A

ADL/IADL Scales

This appendix shows the typical scales used for measuring the level of independence in the Activities of Daily Living (ADL) (Section A.1) and Instrumental Activities of Daily Living (IADL) (Section A.2) execution. They are widely used in the medical field for measuring the functional status of a person.

A.1 Activities of Daily Living Scale

Katz Index of Independence in Activities of Daily Living

ACTIVITIES Points (1 or 0)	INDEPENDENCE: (1 POINT) **NO** supervision, direction or personal assistance	DEPENDENCE: (0 POINTS) **WITH** supervision, direction, personal assistance or total care
BATHING Points: _____	(1 POINT) Bathes self completely or needs help in bathing only a single part of the body such as the back, genital area or disabled extremity.	(0 POINTS) Needs help with bathing more than one part of the body, getting in or out of the tub or shower. Requires total bathing.
DRESSING Points: _____	(1 POINT) Gets clothes from closets and drawers and puts on clothes and outer garments complete with fasteners. May have help tying shoes.	(0 POINTS) Needs help with dressing self or needs to be completely dressed.
TOILETING Points: _____	(1 POINT) Goes to toilet, gets on and off, arranges clothes, cleans genital area without help.	(0 POINTS) Needs help transferring to the toilet, cleaning self or uses bedpan or commode.
TRANSFERRING Points: _____	(1 POINT) Moves in and out of bed or chair unassisted. Mechanical transferring aides are acceptable.	(0 POINTS) Needs help in moving from bed to chair or requires a complete transfer.
CONTINENCE Points: _____	(1 POINT) Exercises complete self control over urination and defecation.	(0 POINTS) Is partially or totally incontinent of bowel or bladder.
FEEDING Points: _____	(1 POINT) Gets food from plate into mouth without help. Preparation of food may be done by another person.	(0 POINTS) Needs partial or total help with feeding or requires parenteral feeding.

TOTAL POINTS = _____ *6 = High (patient independent)* *0 = Low (patient very dependent)*

Slightly adapted from Katz S., Down, T.D., Cash, H.R. et al. (1970) Progress in the Development of the Index of ADL. *Gerontologist*

***Figure A.1**: ADL scale*

A.2 Instrumental Activities of Daily Living Scale

INSTRUMENTAL ACTIVITIES OF DAILY LIVING SCALE (IADL)
M.P. Lawton & E.M. Brody

A. Ability to use telephone

1. Operates telephone on own initiative; looks up and dials numbers, etc.	1
2. Dials a few well-known numbers	1
3. Answers telephone but does not dial	1
4. Does not use telephone at all.	0

B. Shopping

1. Takes care of all shopping needs independently	1
2. Shops independently for small purchases	0
3. Needs to be accompanied on any shopping trip.	0
4. Completely unable to shop.	0

C. Food Preparation

1. Plans, prepares and serves adequate meals independently	1
2. Prepares adequate meals if supplied with ingredients	0
3. Heats, serves and prepares meals or prepares meals but does not maintain adequate diet.	0
4. Needs to have meals prepared and served.	0

D. Housekeeping

1. Maintains house alone or with occasional assistance (e.g. "heavy work domestic help")	1
2. Performs light daily tasks such as dish-washing, bed making	1
3. Performs light daily tasks but cannot maintain acceptable level of cleanliness.	1
4. Needs help with all home maintenance tasks.	1
5. Does not participate in any housekeeping tasks.	0

E. Laundry

1. Does personal laundry completely	1
2. Launders small items; rinses stockings, etc.	1
3. All laundry must be done by others.	0

F. Mode of Transportation

1. Travels independently on public transportation or drives own car.	1
2. Arranges own travel via taxi, but does not otherwise use public transportation.	1
3. Travels on public transportation when accompanied by another.	1
4. Travel limited to taxi or automobile with assistance of another.	0
5. Does not travel at all.	0

G. Responsibility for own medications

1. Is responsible for taking medication in correct dosages at correct time.	1
2. Takes responsibility if medication is prepared in advance in separate dosage.	0
3. Is not capable of dispensing own medication.	0

H. Ability to Handle Finances

1. Manages financial matters independently (budgets, writes checks, pays rent, bills goes to bank), collects and keeps track of income.	1
2. Manages day-to-day purchases, but needs help with banking, major purchases, etc.	1
3. Incapable if handling money.	0

Source: Lawton, M.P., and Brody, E.M. "Assessment of older people: Self-maintaining and instrumental activities of daily living." Gerontologist 9:179-186, (1969).

***Figure A.2**: IADL scale*

List of Figures

List of Tables

Bibliography

[AAL] Ambient assisted living. Online http://www.aal-europe.eu/. *cited on pp.* 2

[Abowd *et al.* 1996] Gregory D. Abowd, Christopher G. Atkenson, Ami Feinstein, Cindy Hmelo, Rob Kooper, Sue Long, Nitin Nick Sawhney, and Mikiya Tani. Teaching and Learning as Multimedia Authoring: The Classroom 2000 Project. In *Proceedings of the 4th ACM Conference on Multimedia (Multimedia'96)*, pages 187–198, Boston, MA, USA, November 1996. *cited on pp.* 1

[Abowd *et al.* 1997] Gregory D. Abowd, Anind K. Dey, Robert Orr, and Jason Brotherton. Context-Awareness in Wearable and Ubiquitous Computing. In *Proceedings of the 1st International Symposium on Wearable Computers (ISWC'97)*, pages 179–180, Cambridge, MA, USA, October 1997. *cited on pp.* 9

[Abowd *et al.* 2000] Gregory D. Abowd, Christopher G. Atkeson, Aaron F. Bobick, Irfan A. Essa, Blair MacIntyre, Elizabeth D. Mynatt, and Thad E. Starner. Living Laboratories: The Future Computing Environments Group at the Georgia Institute of Technology. In *Extended Abstracts of the ACM conference on Human Factors in Computing Systems (CHI'00)*, pages 215–216, The Hague, The Netherlands, April 2000. *cited on pp.* 10

[Aghajan *et al.* 2007] Hamid Aghajan, Juan Carlos Augusto, Chen Wu, Paul McCullagh, and Julie-Ann Walkden. Distributed Vision-Based Accident Management for Assisted Living. In *Proceedings of the 5th International Conference on Smart Homes and Health Telematics (ICOST'07)*, pages 196–205, Nara, Japan, June 2007. *cited on pp.* 12

[Albaina *et al.* 2009] Inaki Merino Albaina, Thomas Visser, Charles A.P.G. van der Mast, and Martijn H. Vastenburg. Flowie: A Persuasive Virtual Coach to Motivate Elderly Individuals to Walk. In *Proceedings of the 3rd International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2009)*, London, UK, April 2009. *cited on pp.* 11

[Allin *et al.* 2003] S.J. Allin, A. Bharucha, J. Zimmerman, D. Wilson, M.J. Roberson, S. Stevens, H. Wactlar, and C.G. Atkeson. Toward the Automatic Assessment of Behavioral Disturbances of Dementia. In *Proceedings of the 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications (UbiHealth'03)*, Seattle, WA, USA, October 2003. *cited on pp.* 11

[Amft *et al.* 2007] Oliver Amft, Martin Kusserow, and Gerhard Tröster. Probabilistic parsing of dietary activity events. In *Proceedings of the 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN'07)*, pages 242–247, Aachen, Germany, March 2007. *cited on pp.* 10, 11

[Anderson and Moore 2005] Brigham Anderson and Andrew Moore. Active Learning for Hidden Markov Models: Objective Functions and Algorithms. In *Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*, pages 9–16, Bonn, Germany, August 2005. *cited on pp.* 17

[Andrews *et al.* 2003] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support Vector Machines for Multiple-Instance Learning. In *Proceedings of the 17th Annual Conference On Neural Information Processing Systems (NIPS'03)*, pages 561–568, Vancuver, BC, Canada, December 2003. *cited on pp.* 67

[Antifakos *et al.* 2002] Stavros Antifakos, Florian Michahelles, and Bernt Schiele. Proactive Instructions for Furniture Assembly. In *Proceedings of the 4th International Conference on Ubiquitous Computing (UbiComp'02)*, pages 351–360, Göteborg, Sweden, September 2002. *cited on pp.* 10

[Ashbrook and Starner 2003] Daniel Ashbrook and Thad Starner. Learning Significant Locations and Predicting User Movement with GPS. In *Proceedings of the 6th International Symposium on Wearable Computers (ISWC'02)*, pages 101–108, Seattle, WA, USA, October 2003. *cited on pp.* 12

[Backman *et al.* 2006] Anders Backman, Kenneth Bodin, Gösta Bucht, Lars-Erik Janlert, Marcus Maxhall, Thomas Pederson, Daniel Sjölie, Björn Sondell, and Dipak Surie. EasyADL - Wearable Support System for Independent Life Despite Dementia. In *Proceedings of the CHI 2006 Workshop on Designing Technology for People with Cognitive Impairments*, Montreal, Canada, April 2006. *cited on pp.* 11

[Balcan *et al.* 2005] Maria-Florina Balcan, Avrim Blum, Patrick Pakyan Choi, John Lafferty, Brian Pantano, Mugizi Robert Rwebangira, and Xiaojin Zhu. Person Identification in Webcam Images: An Application of Semi-Supervised Learning. In *Proceedings of the ICML'05 Workshop on Learning with Partially Classified Training Data*, Bonn, Germany, August 2005. *cited on pp.* 87

[Bao and Intille 2004] Ling Bao and Stephen S. Intille. Activity Recognition from User-Annotated Acceleration Data. In *Proceedings of the 2nd International Conference on Pervasive Computing (Pervasive'04)*, pages 1–17, Vienna, Austria, April 2004. *cited on pp.* 9, 10, 12, 13, 14, 16

[Bardram and Christensen 2007] Jakob E. Bardram and Henrik B. Christensen. Pervasive Computing Support for Hospitals: An Overview of the Activity-Based Computing Project. *IEEE Pervasive Computing*, 6(1):44–51, 2007. *cited on pp.* 1

[Beaudin *et al.* 2007] Jennifer S. Beaudin, Stephen S. Intille, Emmanuel Munguia Tapia, Randy Rockinson, and Margeret E. Morris. Context-Sensitive Microlearning of Foreign Language Vocabulary on a Mobile Device. In *Proceedings of the European Con-*

ference on Ambient Intelligence (AmI'07), pages 55–72, Darmstadt, Germany, November 2007. *cited on pp.* 1

[Blanke and Schiele 2009] Ulf Blanke and Bernt Schiele. Daily Routine Recognition through Activity Spotting. In *Proceedings of the 4th International Symposium on Location and Context Awareness (LoCA'09)*, Tokyo, Japan, May 2009. *cited on pp.* 102

[Blum and Mitchell 1998] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98)*, pages 92–100, Madison, WI, USA, July 1998. *cited on pp.* 18, 52, 56, 58

[Brashear *et al.* 2003] Helene Brashear, Thad Starner, Paul Lukowicz, and Holger Junker. Using Multiple Sensors for Mobile Sign Language Recognition. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC'03)*, pages 45–52, White Plains, NY, USA, October 2003. *cited on pp.* 13

[Brashear *et al.* 2006] Helene Brashear, Valerie Henderson, Kwang-Hyun Park, Harley Hamilton, Seungyon Lee, and Thad Starner. American Sign Language Recognition in Game Development for Deaf Children. In *Proceedings of the 8th International Conference on Computers and Accessibility (ASSETS'06)*, pages 79–86, Portland, OR, USA, October 2006. *cited on pp.* 1

[Bulling *et al.* 2008] Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. Robust Recognition of Reading Activity in Transit Using Wearable Electrooculography. In *Proceedings of the International Conference on Pervasive Computing (Pervasive'08)*, pages 19–37, Sydney, Australia, May 2008. *cited on pp.* 13

[Burges 1998] Christopher J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998. *cited on pp.* 27

[Chambers *et al.* 2002] Graeme S. Chambers, Svetha Venkatesh, Geoff A.W. West, and Hung H. Bui. Hierarchical Recognition of Intentional Human Gestures for Sports Video Annotation. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02)*, pages 1082–1085, Quebec, Canada, August 2002. *cited on pp.* 10

[Chan *et al.* 2003] H.L. Chan, C.H. Lin, and Y.L. Ko. Segmentation of Heart Rate Variability in Different Physical Activities. In *Proceedings of the Computers in Cardiology*, pages 97–100, Thessaloniki, Greece, September 2003. *cited on pp.* 13

[Chang *et al.* 2007] Keng Hao Chang, Mike Y. Chen, and John Canny. Tracking Free-Weight Exercises. In *Proceedings of the 9th International Conference on Ubiquitous Computing (UbiComp'07)*, pages 19–37, Innsbruck, Austria, September 2007. *cited on pp.* 1, 10

[Chang *et al.* 2008] Yao-Jen Chang, Chien-Nien Chen, Li-Der Chou, and Tsen-Yung Wang. A Novel Indoor Wayfinding System Based on Passive RFID for Individuals with Cognitive Impairments. In *Proceedings of the 2nd International Conference on*

Pervasive Computing Technologies for Healthcare (Pervasive Health 2008), Tampere, Finland, January 2008. *cited on pp.* 11

[Chapelle *et al.* 2006] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006. *cited on pp.* 17, 52, 56

[Chen *et al.* 2005] Jianfeng Chen, Jianmin Zhang, Alvin Harvey Kam, and Louis Shue. An Automatic Acoustic Bathroom Monitoring System. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS'05)*, volume 2, pages 1750–1753, Kobe, Japan, May 2005. *cited on pp.* 10, 12

[Cheng and Lukowicz 2008] Jingyuan Cheng and Paul Lukowicz. Towards Wearable Capacitive Sensing of Physiological Parameters. In *Proceedings of the 2nd International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2008)*, pages 272–273, Tampere, Finland, January 2008. *cited on pp.* 13

[Choudhury and Pentland 2003] Tanzeem Choudhury and Alex Pentland. Sensing and Modeling Human Networks using the Sociometer. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC'03)*, pages 216–222, White Plains, NY, USA, October 2003. *cited on pp.* 12

[Choudhury *et al.* 2006] Tanzeem Choudhury, Matthai Philipose, Danny Wyatt, and Jonathan Lester. Towards Activity Databases: Using Sensors and Statistical Models to Summarize Peoples Lives. *IEEE Data Engineering*, 29(1):49–58, 2006. *cited on pp.* 11

[Choudhury *et al.* 2008] Tanzeem Choudhury, Gaetano Borriello, Sunny Consolvo, Dirk Haehnel, Beverly Harrison, Bruce Hemingway, Jeff Hightower, Pedja Klasnja, Karl Koscher, Anthony LaMarca, Jonathan Lester, James A. Landay, Louis Legrand, Ali Rahimi ad Adam Rea, and Danny Wyatt. The Mobile Sensing Platform: An Embedded Activity Recognition System. *IEEE Pervasive Computing*, 7(2):32–41, April–June 2008. *cited on pp.* 13

[Clarkson and Pentland 1999] Brian Clarkson and Alex Pentland. Unsupervised Clustering of Ambulatory Audio and Video. In *Proceedings of the IEEE Internetional Conference of Acoustics, Speech, and Signal Processing (ICASSP'99)*, pages 3037–3040, Phoenix, AZ, USA, March 1999. *cited on pp.* 10, 13, 17

[Consolvo and Walker 2003] Sunny Consolvo and Miriam Walker. Using the Experience Sampling Method to Evaluate Ubicomp Applications. *IEEE Pervasive Computing*, 2(2):24–31, 2003. *cited on pp.* 15

[Consolvo *et al.* 2008a] Sunny Consolvo, Predrag Klasnja, David W. McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A. Landay. Flowers or a Robot Army? Encouraging Awareness and Activity with Personal, Mobile Displays. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp'08)*, pages 54–63, Seoul, Korea, September 2008. *cited on pp.* 11

[Consolvo *et al.* 2008b] Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis

LeGrand, Ryan Libby, Ian Smith, and James A. Landay. Activity Sensing in the Wild: A Field Trial of UbiFit Garden. In *Proceedings of the 26th Annual Conference on Human Factors in Computing Systems (CHI'08)*, pages 1797–1806, Florence, Italy, April 2008. *cited on pp.* 1

[Csikszentmihalyi and Larson 1987] Mihaly Csikszentmihalyi and Reed Larson. Validity and Reliability of the Experience-Sampling Method. *Journal on Nervous and Mental Disease*, 175(9):526–536, September 1987. *cited on pp.* 14

[Cuzzort and Starner 2008] Stephen Cuzzort and Thad Starner. AstroWheelie: A wheelchair based exercise game. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 113–114, Pittsburgh, PA, USA, September 2008. *cited on pp.* 11

[Degen *et al.* 2008] Thomas Degen, Heinz Jaeckel, Michael Rufer, and Stefan Wyss. Speedy: A Fall Detector in a Wrist Watch. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC'03)*, pages 184–188, White Plains, NY, USA, October 2008. *cited on pp.* 11

[Dey 2000] Anind K. Dey. *Providing Architectural Support for Building Context-Aware Applications*. PhD thesis, Georgia Institute of Technology, 2000. *cited on pp.* 1

[Doukas and Maglogiannis 2008] Charalampos Doukas and Ilias Maglogiannis. Advanced Patient or Elder Fall Detection based on Movement and Sound Data. In *Proceedings of the 2nd International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2008)*, Tampere, Finland, January 2008. *cited on pp.* 11

[Du *et al.* 2008] Kejun Du, Daqin Zhang, Xingshe Zhou, Mounir Mokhtari, Mossab Hariz, and Weijun Qin. HYCARE: A Hybrid Context-Aware Reminding Framework for Elders with Mild Dementia. In *Proceedings of the 6th International Conference on Smart Homes and Health Telematics (ICOST'08)*, pages 9–17, Ames, IA, USA, June 2008. *cited on pp.* 11

[Duong *et al.* 2005] Thi V. Duong, Hung H. Bui, Dinh Q. Phung, and Svetha Venkatesh. Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 838–845, San Diego, CA, USA, June 2005. *cited on pp.* 10, 12

[Ermes *et al.* 2008] Miikka Ermes, Juha Pärkkä, Jani Mäntyjärvi, and Ilkka Korhonen. Detection of Daily Activities and Sports With Wearable Sensors in Controlled and Uncontrolled Conditions. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):20–26, January 2008. *cited on pp.* 10

[Farringdon *et al.* 1999] Jonny Farringdon, Andrew J. Moore, Nancy Tilbury, James Church, and Pieter D. Biemond. Wearable Sensor Badge And Sensor Jacker for Context Awareness. In *Proceedings of the 3rd International Symposium on Wearable Computers (ISWC'99)*, pages 107–113, San Francisco, CA, USA, Octeber 1999. *cited on pp.* 9

[Fishkin *et al.* 2005] Kenneth P. Fishkin, Matthai Philipose, and Adam Rea. Hands-on RFID: Wireless Wearables for Detecting Use of Objects. In *Proceedings of the 9th IEEE International Symposium on Wearable Computers (ISWC'05)*, pages 38–41, Osaka, Japan, October 2005. *cited on pp.* 5, 33

[French *et al.* 2008] Brian French, Divya Tyamagundlu, Daniel P. Siewiorek, Asim Smailagic, and Dan Ding. Towards a Virtual Coach for manual wheelchair users. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 77–80, Pittsburgh, PA, USA, September 2008. *cited on pp.* 11

[Froehlich *et al.* 2007] Jon Froehlich, Mike Y. Chen, Sunny Consolvo, Beverly Harrison, and James A. Landay. MyExperience: A System for In situ Tracing and Capturing of User Feedback on Mobile Phones. In *Proceedings of MobiSys*, pages 57–70, San Juan, Puerto Rico, June 2007. *cited on pp.* 14, 15

[Gao *et al.* 2004] Jiang Gao, Alexander G. Hauptmann, Ashok Bharucha, and Howard D. Wactlar. Dining Activity Analysis Using a Hidden Markov Model. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, Cambridge, United Kingdom, August 2004. *cited on pp.* 10, 12

[Gavrila 1999] Dariu M. Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding (CVIU)*, 73(1):82–98, January 1999. *cited on pp.* 12

[Gehler and Chapelle 2007] Peter V. Gehler and Olivier Chapelle. Deterministic Annealing for Multiple-Instance Learning. In *Proceedings of the 11th International Conference of Artificial Intelligence and Statistics (AISTATS'07)*, pages 123–130, San Juan Puerto Rico, March 2007. *cited on pp.* 102

[Grabner and Bischof 2006] Helmut Grabner and Horst Bischof. On-line Boosting and Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 260–267, New York, NY, USA, June 2006. *cited on pp.* 103

[Guan *et al.* 2007] Donghai Guan, Weiwei Yuan, Young-Koo Lee, Andrey Gavrilov, and Sungyoung Lee. Activity Recognition Based on Semi-supervised Learning. In *Proceedings of the 13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA'07)*, pages 469–475, Daegu, Korea, August 2007. *cited on pp.* 17, 52, 56

[Hanser *et al.* 2008] Friedrich Hanser, Agnes Gruenerbl, Clemens Rodegast, and Paul Lukowicz. Design and Real Life Deployment of a Pervasive Monitoring System for Dementia Patients. In *Proceedings of the 2nd International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2008)*, Tampere, Finland, January 2008. *cited on pp.* 11

[Hektner and Csikszentmihalyi 2002] Joel M. Hektner and Mihaly Csikszentmihalyi. The Experience Sampling Method: Measuring the Context and Content of Lives. *Handbook of Environmental Psychology*, pages 233–243, 2002. *cited on pp.* 14

[Hoey *et al.* 2005] Jesse Hoey, Pascal Poupart, Craig Boutilier, and Alex Mihailidis. Semi-supervised learning of a pomdp model of patientcaregiver interactions. In *Proceedings of the IJCAI Workshop on Modeling Others from Observations (MOO 2005)*, pages 101–110, Edinburgh, Scotland, August 2005. *cited on pp.* 17

[Hoey *et al.* 2007] Jesse Hoey, Alex von Bertoldi, Pascal Poupart, and Alex Mihailidis. Assisting Persons with Dementia during Handwashing Using a Partially Observable Markov Decision Process. In *Proceedings of the 5th International Conference on Computer Vision Systens (ICVS'07)*, Bielefeld University, Germany, March 2007. *cited on pp.* 10, 11, 104

[Hu *et al.* 2008] Derek Hao Hu, Sinno Jialin Pan, Vincent Wenchen Zheng, Nathan Nan Liu, and Qiang Yang. Real World Activity Recognition with Multiple Goals. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 30–39, Seoul, South Korea, September 2008. *cited on pp.* 102

[Huang *et al.* 2008] Kevin Huang, Ellen Yi-Luen Do, and Thad Starner. PianoTouch: A Wearable Haptic Piano Instruction System For Passive Learning of Piano Skills. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 41–44, Pittsburgh, PA, USA, September 2008. *cited on pp.* 1

[Huỳnh and Schiele 2006a] Tâm Huỳnh and Bernt Schiele. Towards Less Supervision in Activity Recognition from Wearable Sensors. In *Proceedings of the 10th IEEE International Symposium on Wearable Computers (ISWC'06)*, pages 19–37, Montreux, Switzerland, October 2006. *cited on pp.* 16

[Huỳnh and Schiele 2006b] Tâm Huỳnh and Bernt Schiele. Unsupervised Discovery of Structure in Activity Data using Multiple Eigenspaces. In *Proceedings of the 2nd International Workshop on Location- and Context-Awareness (LoCA'06)*, pages 151–167, Dublin, Ireland, May 2006. *cited on pp.* 10, 17

[Huỳnh *et al.* 2007] Tâm Huỳnh, Ulf Blanke, and Bernt Schiele. Scalable Recognition of Daily Activities with Wearable Sensors. In *Proceedings of the 3rd International Symposium On Location- and Context- Awareness (LoCA'07)*, pages 50–67, Oberpfaffenhofen, Germany, September 2007. *cited on pp.* 10, 13, 16, 70

[Huỳnh *et al.* 2008] Tâm Huỳnh, Mario Fritz, and Bernt Schiele. Discovery of Activity Patterns using Topic Models. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp'08)*, pages 10–19, Seoul, Korea, September 2008. *cited on pp.* 5, 10, 14, 17, 23, 71, 72, 102, 103

[Intille *et al.* 2003] Stephen S. Intille, John Rondoni, Charles Kukla, Isabel Ancona, and Ling Bao. A Context-Aware Experience Sampling Tool. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'03)*, pages 972–973, Fort Lauderdale, FL, USA, 2003. *cited on pp.* 15, 102

[Intille *et al.* 2006] Stephen S. Intille, Kent Larson, Emmanuel Munguia Tapia, Jennifer S. Beaudin, Pallavi Kaushik, Jason Nawyn, and Randy Rockinson. Using a Live-In Laboratory for Ubiquitous Computing Research. In *Proceedings of the 4th Inter-*

national Conference on Pervasive Computing (Pervasive'06), pages 349–365, Dublin, Ireland, May 2006. *cited on pp.* 10, 14, 22

[Jafari *et al.* 2007] Roozbeh Jafari, Wenchao Li, Ruzena Bajcsy, Steven Glaser, and Shankar Sastry. Physical Activity Monitoring for Assisted Living at Home. In *Proceedings of the 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN'07)*, pages 213–219, Aachen, Germany, March 2007. *cited on pp.* 1, 11

[Joachims 1999] Thorsten Joachims. Making Large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, pages 41–56, 1999. *cited on pp.* 27

[Junker *et al.* 2004] Holger Junker, Paul Lukowicz, and Jani Mäntyjarvi. Proceedings of the pervasive 2004 workshop on benchmarks and a database for context recognition. Zurich, Switzerland, April 2004. *cited on pp.* 5

[Kapoor and Horvitz 2007] Ashish Kapoor and Eric Horvitz. On Discarding, Caching, and Recalling Samples in Active Learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI'07)*, Vancuver, BC, Canada, July 2007. *cited on pp.* 18

[Kapoor and Horvitz 2008] Ashish Kapoor and Eric Horvitz. Experience Sampling for Building Predictive User Models: A Comparative Study. In *Proceedings of the 26th Conference on Human Factors in Computing Systems (CHI'08)*, pages 657–666, Florence, Italy, April 2008. *cited on pp.* 18, 60

[Katz 1983] Sidney Katz. Assessing Self-Maintenance: Activities of Daily Living, Mobility, and Instrumental Activities of Daily Living. *Journal of the American Geriatrics Society*, 31(12):721–726, 1983. *cited on pp.* 2

[Kern *et al.* 2003] Nicky Kern, Bernt Schiele, and Albrecht Schmidt. Multi-Sensor Activity Context Detection for Wearable Computing. In *Proceedings of the 1st European Symposium on Ambient Intelligence (EUSAI'03)*, pages 220–232, Veldhoven, The Netherlands, November 2003. *cited on pp.* 9

[Kern *et al.* 2004] Nicky Kern, Stavros Antifakos, Bernt Schiele, and Adrian Schwaninger. A Model for Human Interruptability: Experimental Evaluation and Automatic Estimation from Wearable Sensors. In *Proceedings of the 8th International Symposium on Wearable Computers (ISWC'04)*, pages 158–165, Washington, DC, USA, November 2004. *cited on pp.* 13

[Klasnja *et al.* 2008] Predrag Klasnja, Beverly L. Harrison, Louis LeGrand, Anthony LaMarca, Jon Froehlich, and Scott E. Hudson. Using Wearable Sensors and Real Time Inference to Understand Human Recall of Routine Activities. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp'08)*, pages 154–163, Seoul, Korea, September 2008. *cited on pp.* 15

[Krause *et al.* 2003] Andreas Krause, Daniel P. Siewiorek, Asim Smailagic, and Jonny Farringdon. Unsupervised, Dynamic Identification of Physiological and Activity Context in Wearable Computing. In *Proceedings of the 7th International Symposium on*

Wearable Computers (ISWC'03), pages 88–97, White Plaing, NY, USA, October 2003. *cited on pp.* 17

[Kunze *et al.* 2006] Kai Kunze, Michael Barry, Ernst A. Heinz, Paul Lukowicz, Dennis Majoe, and Jürg Gutknecht. Towards Recognizing Tai Chi–An Initial Experiment Using Wearable Sensors. In *Proceedings of the 3rd International Forum on Applied Wearable Computing (IFAWC'06)*, Bremen, Germany, March 2006. *cited on pp.* 1, 10

[Lackovic *et al.* 2000] Igor Lackovic, Vedran Bilas, Ante Santic, and Vasilije Nikolic. Measurement of gait parameters from free moving objects. *Measurement*, 27(2):121–131, 2000. *cited on pp.* 11

[Lee and Dey 2008] Matthew L. Lee and Anind K. Dey. Lifelogging Memory Appliance for People with Episodic Memory Impairment. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp'08)*, pages 44–53, Seoul, Korea, September 2008. *cited on pp.* 11

[Lee and Mase 2002] Seon-Woo Lee and Kenji Mase. Activity and Location Recognition Using Wearable Sensors. *IEEE Pervasive Computing*, 1(2):24–32, July–September 2002. *cited on pp.* 9

[Lester *et al.* 2005] Jonathan Lester, Tanzeem Choudhury, Nicky Kern, Gaetano Borriello, and Blake Hannford. A Hybrid Discriminative/Generative Approach for Modeling Human Activities. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)*, pages 766–772, Edinburg, Scotland, August 2005. *cited on pp.* 16, 39

[Lester *et al.* 2006] Jonathan Lester, Tanzeem Choudhury, and Gaetano Borriello. A Practical Approach to Recognizing Physical Activities. In *Proceedings of the 4th International Conference on Pervasive Computing (Pervasive'06)*, pages 1–16, Dublin, Ireland, May 2006. *cited on pp.* 9, 13, 16

[Liao *et al.* 2005] Lin Liao, Dieter Fox, and Henry Kautz. Location Based Activity Recognition. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NIPS'05)*, pages 787–794, Vancuver, B.C., Canada, December 2005. *cited on pp.* 12

[Liu *et al.* 2009a] Alan L. Liu, Harlan Hile, Gaetano Borriello, Henry Kautz, Pat A. Brown, Mark Harniss, and Kurt Johnson. Informing the Design of an Automated Wayfinding System for Individuals with Cognitive Impairments. In *Proceedings of the 3nd International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2009)*, London, UK, April 2009. *cited on pp.* 11

[Liu *et al.* 2009b] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory Under-Sampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man and Cybernetics - Part B*, 39(2):539–550, 2009. *cited on pp.* 102

[Logan *et al.* 2007] Beth Logan, Jennifer Healey, Matthai Philipose, Emmanuel Munguia Tapia, and Stephen Intille. A Long-Term Evaluation of Sensing Modalities for Activity Recognition. In *Proceedings of the Ubicomp '07*, pages 483–500, Innsbruck, Austria, September 2007. Dataset http://architecture.

mit.edu/house_n/data/PlaceLab/PlaceLab.htm. *cited on pp.* 5, 10, 12, 14, 16, 18, 22, 29, 52, 53, 54, 72

[Lukowicz *et al.* 2004] Paul Lukowicz, Jamie A. Ward, Holger Junker, Mathias Stäger, Gerhard Tröster, Amin Atrash, and Thad Starner. Recognizing Workshop Activity Using Body Worn Microphones and Accelerometers. In *Proceedings of the 2nd International Conference on Pervasive Computing (Pervasive'04)*, pages 18–32, Vienna, Austria, April 2004. *cited on pp.* 10, 16

[Lukowicz *et al.* 2006] Paul Lukowicz, Friedrich Hanser, Cristoph Szubski, and Wolfgang Schobersberger. Detecting and Interpreting Muscle Activity with Wearable Force Sensors. In *Proceedings of the 4th International Conference on Pervasive Computing (Pervasive'06)*, pages 101–116, Dublin, Ireland, May 2006. *cited on pp.* 13

[Lukowicz *et al.* 2007] Paul Lukowicz, Andreas Timm-Giel, Michael Lawo, and Otthein Herzog. WearIT@work: Toward Real-World Industrial Wearable Computing. *IEEE Pervasive Computing*, 6(4):8–13, October–December 2007. *cited on pp.* 1, 10

[Mahdaviani and Choudhury 2007] Maryam Mahdaviani and Tanzeem Choudhury. Fast and Scalable Training of Semi-Supervised CRFs with Application to Activity Recognition. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS'07)*, Vancuver, BC, Canada, December 2007. *cited on pp.* 17

[Mäntyjärvi *et al.* 2001] Jani Mäntyjärvi, Johan Himberg, and Tapio Seppänen. Recognizing Human Motion With Multiple Acceleration Sensors. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 747–752, Tuscon, AZ, USA, October 2001. *cited on pp.* 9

[Maurer *et al.* 2006] Uwe Maurer, Asim Smailagic, Daniel P. Siewiorek, and Michael Deisher. Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions. In *Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*, pages 113–116, Boston, MA, USA, April 2006. *cited on pp.* 9, 12, 14, 16

[McCallum and Nigam 1998] Andrew McCallum and Kamal Nigam. A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of the AAAI'98 Workshop on Learning for Text Categorization*, pages 41–48, Madison WI, USA, July 1998. *cited on pp.* 25

[Meyer *et al.* 2008] F. Meyer, G. von Bögel, C. Ressel, and T. Dimitrov. inHaus2: Intelligent construction site logistics. In *Proceedings of the European Workshop on RFID Systems and Technologies (RFID SysTech'08)*, Freiburg, Germany, June 2008. *cited on pp.* 10

[Mihailidis *et al.* 2003] A. Mihailidis, G. Fernie, and W. Cleghorn. The development of computerized cueing device to help people with dementia to be more independent. *Technology and Disability*, 2(2):173–189, 2003. *cited on pp.* 11

[Mihailidis *et al.* 2004] Alex Mihailidis, Brent Carmichael, and Jennifer Boger. The Use of Computer Vision in an Intelligent Environment to Support Aging-in-Place, Safety,

and Independence in the Home. *IEEE Transactions on Information Technology in Biomedicine*, 8(3):238–247, September 2004. *cited on pp.* 12

[Minnen *et al.* 2005] David Minnen, Thad Starner, Jamie A. Ward, Paul Lukowicz, and Gerhard Tröster. Recognizing and Discovering Human Actions from On-body Sensor Data. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'05)*, Amsterdam, The Netherlands, July 2005. *cited on pp.* 13

[Minnen *et al.* 2006a] David Minnen, Thad Starner, Irfan Essa, and Charles Isbell. Discovering Characteristic Actions from On-Body Sensor Data. In *Proceedings of the 10th IEEE International Symposium on Wearable Computers (ISWC'06)*, pages 11–18, Montreux, Switzerland, October 2006. *cited on pp.* 10, 17, 102

[Minnen *et al.* 2006b] David Minnen, Tracy Westeyn, Thad Starner, Jamie A. Ward, and Paul Lukowicz. Performance Metrics and Evaluation Issues for Continuous Activity Recognition. In *Proceedings of the Performance Metrics for Intelligent Systems Workshop (PerMIS'06)*, Gaithersburg, MD, USA, August 2006. *cited on pp.* 5

[Minnen *et al.* 2007] David Minnen, Tracy Westeyn, Daniel Ashbrook, Peter Presti, and Thad Starner. Recognizing Soldier Activities in the Field. In *Proceedings of the 4th International Workshop on Wearable and Implantable Body Sensor Networks*, Aachen, Germany, March 2007. *cited on pp.* 10, 16

[Modayil *et al.* 2008] Joseph Modayil, Tongxin Bai, and Henry Kautz. Improving the Recognition of Interleaved Activities. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 40–43, Seoul, South Korea, September 2008. *cited on pp.* 102

[Morris *et al.* 2003] Margaret Morris, Jay Lundell, Eric Dishman, and Brad Needham. New Perspectives on Ubiquitous Computing from Ethnographic Study of Elders with Cognitive Decline. In *Proceedings of the 5th International Conference on Ubiquitous Computing (UbiComp'03)*, pages 227–242, Seattle, WA, USA, October 2003. *cited on pp.* 2

[Morris *et al.* 2005] Margaret Morris, Stephen S. Intille, and Jennifer S. Beaudin. Embedded Assessment: Overcoming Barriers to Early Detection with Pervasive Computing. In *Proceedings of the 3rd International Conference on Pervasive Computing (Pervasive'05)*, pages 333–346, Munich, Germany, May 2005. *cited on pp.* 2

[Mozer 1998] Michael C. Mozer. The Neural Network House: An Environment that Adapts to its Inhabitants. In *Proceedings of the American Association for Artificial Intelligence Spring Symposium on Intelligent Environments*, pages 110–114, Menlo Park, CA, USA, March 1998. *cited on pp.* 10

[Murphy 1998] Kevin Murphy, 1998. HMM Toolbox for Matlab. Online http://www.cs.ubc.ac/ murphyk/Software/HMM/hmm.html. *cited on pp.* 25

[Muslea *et al.* 2000] Ion Muslea, Steven Minton, and Craig A. Knoblock. Selective Sampling with Redundant Views. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI'00)*, pages 621–626, Austin, TX, USA, August 2000. *cited on pp.* 51

[Najafi *et al.* 2003] Bijan Najafi, Kamiar Aminian, Anisoara Paraschiv-Ionescu, Francois Loew, and Cristophe J. Büla. Ambulatory System for Human Motion Analysis Using a Kinematic Sensor: Monitoring of Daily Physical Activity in the Edlerly. *IEEE Transactions on Biomedical Engineering*, 50(6):711–723, June 2003. *cited on pp.* 12

[Nowozin *et al.* 2007] Sebastian Nowozin, Gökhan Bakir, and Koji Tsuda. Discriminative Subsequence Mining for Activity Classification. In *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, October 2007. *cited on pp.* 12

[Ogris *et al.* 2005] Georg Ogris, Thomas Stiefmeier, Holger Junker, Paul Lukowicz, and Gerhard Tröster. Using Ultrasonic Hand Tracking to Augment Motion Analysis Based Recognition of Manipulative Gestures. In *Proceedings of the 9th IEEE International Symposium on Wearable Computers (ISWC'05)*, pages 152–159, Osaka, Japan, October 2005. *cited on pp.* 12

[Ogris *et al.* 2008] Georg Ogris, Thomas Stiefmeier, Paul Lukowicz, and Gerhard Tröster. Using a complex Multi-modal On-body Sensor System for Activity Spotting. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 55–62, Pittsburgh, PA, USA, October 2008. *cited on pp.* 10

[Oliver and Flores-Mangas 2007] Nuria Oliver and Fernando Flores-Mangas. HealthGear: Automatic Sleep Apnea Detection and Monitoring with a Mobile Phone. *Journal of Communications*, 2(2):1–9, March 2007. *cited on pp.* 11

[Oliver and Horvitz 2005] Nuria Oliver and Eric Horvitz. A Comparison of HMMs and Dynamic Bayesian Networks for Recognizing Office Activities. In *Proceedings of the 10th Conference on User Modeling (UM'05)*, pages 199–209, Edinburgh, UK, July 2005. *cited on pp.* 10

[Oliver *et al.* 2002] Nuria Oliver, Eric Horvitz, and Ashutosh Garg. Layered Representations for Human Activity Recognition. In *Proceedings of the 4th IEEE International Conference on Multimedia Interfaces (ICMI'02)*, pages 3–8, Pittsburgh, PA, USA, October 2002. *cited on pp.* 10

[Pärkkä *et al.* 2006] Juha Pärkkä, Miikka Ermes, Panu Korpipää, Jani Mäntyjärvi, Johannes Peltola, and Ilkka Korhonen. Activity Classification Using Realistic Data From Wearable Sensors. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):119–128, January 2006. *cited on pp.* 9, 12

[Partridge and Golle 2008] Kurt Partridge and Philippe Golle. On Using Existing Time-Use Study Data for Ubiquitous Computing Applications. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp'08)*, pages 144–153, Seoul, Korea, September 2008. *cited on pp.* 73, 86, 103

[Patel *et al.* 2008] Shwetak N. Patel, Matthew S. Reynolds, and Gregory D. Abowd. Detecting Human Movements by Differential Air Pressure Sensing in HVAC System Ductwork: An Exploration in Infrastructure Mediated Sensing. In *Proceedings of the 6th International Conference on Pervasive Computing (Pervasive'08)*, pages 1–18, Sydney, Australia, May 2008. *cited on pp.* 12

[Patterson *et al.* 2005] Donald J. Patterson, Dieter Fox, Hery Kautz, and Matthai Philipose. Fine-Grained Activity Recognition by Aggregating Abstract Object Usage. In *Proceedings of the 9th IEEE International Symposium on Wearable Computers (ISWC'05)*, pages 44–51, Osaka, Japan, October 2005. *cited on pp.* 12

[Patterson *et al.* 2008] Donald J. Patterson, Lin Liao, Krzysztof Gajos, Michael Collier, Nik Livic, Katherine Olson, Shiaokai Wang, Dieter Fox, and Henry Kautz. Opportunity Knocks: a System to Provide Cognitive Assistance with Transportation Services. In *Proceedings of the 6th International Conference on Ubiquitous Computing (UbiComp'04)*, pages 433–450, 2008. *cited on pp.* 11

[Philipose *et al.* 2004] Matthai Philipose, Kenneth P. Fishkin, Mike Perkowitz, Donald J. Patterson, Dieter Fox, Henry Kautz, and Dirk Hähnel. Inferring Activities from Interactions with Objects. *IEEE Pervasive Computing*, 3(4):50–57, October 2004. *cited on pp.* 10, 12

[Pirkl *et al.* 2008] Gerald Pirkl, Karl Stockinger, Kai Kunze, and Paul Lukowicz. Adapting Magnetic Resonant Coupling Based Relative Positioning Technology for Wearable Activity Recognition. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 47–54, Pittsburgh, PA, USA, October 2008. *cited on pp.* 12

[Quinlan 1993] Ross J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning, San Mateo, CA, 1993. *cited on pp.* 27

[Rabiner 1989] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989. *cited on pp.* 25, 39

[Randell and Muller 2000] Cliff Randell and Henk Muller. Context Awareness by Analysing Accelerometer Data. In *Proceedings of the 4th International Symposium on Wearable Computers (ISWC'00)*, pages 175–176, Atlanta, GA, USA, October 2000. *cited on pp.* 9

[Ravi *et al.* 2005] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. Activity Recognition from Accelerometer Data. In *Proceedings of the 17th Innovative Applications of Artificial Intelligence Conference (IAAI'05)*, pages 1541–1546, Pittsburgh, PA, USA, July 2005. *cited on pp.* 9, 12, 16

[Reddy *et al.* 2008] Sasank Reddy, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. Determining Transportation Mode on Mobile Phones. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, Pittsburgh, PA, USA, September 2008. *cited on pp.* 103

[Salarian *et al.* 2004] A. Salarian, H. Russmann, F.J.G. Vingerhoets, C. Dehollain, Y. Blanc, P.R. Burkhard, and K. Aminian. Gait Assessment in Parkinson's Disease: Toward an Ambulatory System for Long-Term Monitoring. *IEEE Transactions on Biomedical Engineering*, 51(8), August 2004. *cited on pp.* 11

[Schilit *et al.* 1994] Bill N. Schilit, Norman Adams, and Roy Want. Context-Aware Computing Applications. In *Proceedings of the IEEE Workshop on Mobile Computing Sys-*

tems and Applications (WMCSA'94), pages 85–90, Santa Cruz, CA, USA, December 1994. *cited on pp.* 1, 9

[Schmidt 2002] Albrecht Schmidt. *Ubiquitous Computing - Computing in Context*. PhD thesis, Lancaster University, 2002. *cited on pp.* 9

[Schumm *et al.* 2008] Johannes Schumm, Marc Bächlin, Cornelia Setz, Bert Arnrich, Daniel Roggen, and Gerhard Tröster. Effect of Movements on the Electrodermal Response After a Startle Event. In *Proceedings of the 2nd International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2008)*, pages 315–318, Tampere, Finland, January 2008. *cited on pp.* 13

[Shroff *et al.* 2008] Geeta Shroff, Asim Smailagic, and Daniel P. Siewiorek. Wearable Context-Aware Food Recognition for Calorie Monitoring. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 119–120, Pittsburgh, PA, USA, September 2008. *cited on pp.* 11

[Stäger *et al.* 2004] Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. Implementation and Evaluation of a Low-Power Sound-Based User Activity Recognition System. In *Proceedings of the 8th IEEE International Symposium on Wearable Computers (ISWC'04)*, pages 138–141, Arlington, VA, USA, November 2004. *cited on pp.* 12

[Starner *et al.* 1998] Thad Starner, Bernt Schiele, and Alex Pentland. Visual Contextual Awareness in Wearable Computing. In *Proceedings of the 2nd IEEE International Symposium on Wearable Computers (ISWC'98)*, pages 50–57, Pittsburgh, PA, USA, October 1998. *cited on pp.* 13

[Starner 1999] Thad Starner. *Wearable Computing and Contextual Awareness*. PhD thesis, MIT Media Laboratory, 1999. *cited on pp.* 9

[Stiefmeier *et al.* 2006] Thomas Stiefmeier, Georg Ogris, Holger Junker, Paul Lukowicz, and Gerhard Tröster. Combining Motion Sensors and Ultrasonic Hands Tracking for Continuous Activity Recognition in a Maintenance Scenario. In *Proceedings of the 10th IEEE International Symposium on Wearable Computers (ISWC'06)*, pages 97–104, Montreux, Switzerland, October 2006. *cited on pp.* 10, 13

[Stiefmeier *et al.* 2008] Thomas Stiefmeier, Daniel Roggen, Georg Ogris, Paul Lukowicz, and Gerhard Tröster. Wearable Activity Tracking in Car Manufacturing. *IEEE Pervasive Computing*, 7(2), 2008. *cited on pp.* 16

[Stikic and Schiele 2009] Maja Stikic and Bernt Schiele. Activity Recognition from Sparsely Labeled Data Using Multi-Instance Learning. In *Proceedings of the 4th International Symposium on Location and Context Awareness (LoCA'09)*, pages 156–173, Tokyo, Japan, May 2009. *cited on pp.* 6

[Stikic and Van Laerhoven 2007] Maja Stikic and Kristof Van Laerhoven. Recording Housekeeping Activities with Situated Tags and Wrist-Worn Sensors: Experiment Setup and Issues Encountered. In *Proceedings of the INSS Workshop on Wireless Sensor Networks for Health Care (WSNHC 2007)*, Braunschweig, Germany, June 2007. *cited on pp.* 5, 6, 14

[Stikic *et al.* 2008a] Maja Stikic, Tâm Huỳnh, Kristof Van Laerhoven, and Bernt Schiele. ADL Recognition Based on the Combination of RFID and Accelerometer Sensing. In *Proceedings of the 2nd International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2008)*, pages 258–263, Tampere, Finland, January 2008. Dataset http://www.mis.informatik.tu-darmstadt.de/People/maja/datasetpervasivehealth2008.zip. *cited on pp.* 6, 13, 20

[Stikic *et al.* 2008b] Maja Stikic, Kristof Van Laerhoven, and Bernt Schiele. Exploring Semi-Supervised and Active Learning for Activity Recognition. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 81–88, Pittsburgh, PA, USA, October 2008. *cited on pp.* 6, 17, 72

[Stikic *et al.* 2009] Maja Stikic, Diane Larlus, and Bernt Schiele. Multi-Graph Based Semi-Supervised Learning for Activity Recognition. In *Proceedings of the 13th IEEE International Symposium on Wearable Computers (ISWC'09)*, Linz, Austria, September 2009. *cited on pp.* 6

[Subramanya *et al.* 2006] Amarnag Subramanya, Alvin Raj, Jeff Blimes, and Dieter Fox. Recognizing Activities and Spatial Context Using Wearable Sensors. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI'06)*, pages 211–221, Cambridge, MA, USA, July 2006. *cited on pp.* 13, 17

[Tanaka *et al.* 2004] S. Tanaka, K. Motoi, M. Nogawa, and Yamakoshi K. A New Portable Device for Ambulatory Monitoring of Human Posture and Walking Velocity Using Miniature Accelerometers and Gyroscope. In *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'04)*, pages 2283–2286, San Francisco, CA, USA, September 2004. *cited on pp.* 12

[Tapia *et al.* 2004] Emmanuel Munguia Tapia, Stephen S. Intille, and Kent Larson. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In *Proceedings of the 2nd International Conference on Pervasive Computing (Pervasive'04)*, pages 158–175, Vienna, Austria, April 2004. *cited on pp.* 10, 12, 14, 16

[Tapia *et al.* 2006] Emmanuel Munguia Tapia, Stephen S. Intille, Louis Lopez, and Kent Larson. The Design of a Portable Kit of Wireless Sensors for Naturalistic Data Collection. In *Proceedings of the 4th International Conference on Pervasive Computing (Pervasive'06)*, pages 117–134, Dublin, Ireland, May 2006. *cited on pp.* 22

[Tapia *et al.* 2007a] Emmanuel Munguia Tapia, Stephen S. Intille, William Haskell, Kent Larson, Julie Wright, Abby King, and Robert Friedman. Real-Time Recognition of Physical Activities and Their Intensities Using Wireless Accelerometers and a Heart Rate Monitor. In *Proceedings of the 11th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 37–40, Boston, MA, USA, October 2007. *cited on pp.* 10, 13, 16

[Tapia *et al.* 2007b] Emmanuel Munguia Tapia, Stephen S. Intille, and Kent Larson. Portable Wireless Sensors for Object Usage Sensing in the Home: Challenges and Practicalities. In *Proceedings of the European Conference on Ambient Intelligence (AmI'07)*, pages 19–37, Darmstadt, Germany, November 2007. *cited on pp.* 12

[Torralba *et al.* 2004] Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages 762–769, Washington, DC, USA, July 2004. *cited on pp.* 25, 26, 39, 52, 53

[Toscos *et al.* 2008] Tammy Toscos, Anne Faber, Kay Connelly, and Adity Mutsuddi Upoma. Encouraging Physical Activity in Teens: Can technology help reduce barriers to physical activity in adolescent girls? In *Proceedings of the 2nd International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2008)*, Tampere, Finland, January 2008. *cited on pp.* 11

[Tran and Mynatt 2002] Quan T. Tran and Elizabeth D. Mynatt. Cook's Collage: Two Exploratory Designs. In *Extended Abstracts of CHI 2002: Workshop - Technologies for Families*, Minneapolis, MN, USA, April 2002. *cited on pp.* 10

[Van Laerhoven and Aronsen 2007] Kristof Van Laerhoven and Andre Kvist Aronsen. Memorizing What You Did Last Week: Towards Detailed Actigraphy With A Wearable Sensor. In *Proceedings of the 7th International Workshop on Smart Appliances and Wearable Computing (IWSAWC'07)*, pages 47–52, Toronto, Canada, June 2007. *cited on pp.* 5, 13, 34

[Van Laerhoven and Cakmakci 2000] Kristof Van Laerhoven and Ozan Cakmakci. What Shall We Teach our Pants? In *Proceedings of the 4th International Symposium on Wearable Computers (ISWC'00)*, pages 77–83, Atlanta, GA, USA, October 2000. *cited on pp.* 9

[Van Laerhoven and Gellerson 2004] Kristof Van Laerhoven and Hans-Werner Gellerson. Spine versus Porcupine: a Study in Distributed Wearable Activity Recognition. In *Proceedings of the 8th International IEEE Symposium on Wearable Computers (ISWC'04)*, pages 142–150, Arlington, VA, USA, November 2004. *cited on pp.* 12, 13

[Van Laerhoven *et al.* 2003] Kristof Van Laerhoven, Nicky Kern, Hans-Werner Gellersen, and Bernt Schiele. Towards a Wearable Inertial Sensor Network. In *Proceedings of the IEE Conference Eurowearable'03*, pages 125–130, Birmingham, UK, September 2003. *cited on pp.* 9, 16

[Van Laerhoven *et al.* 2006] Kristof Van Laerhoven, Hans-Werner Gellerson, and Yanni G. Malliaris. Long-Term Activity Monitoring with a Wearable Sensor Node. In *Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*, pages 171–174, Boston, MA, USA, April 2006. *cited on pp.* 11, 12

[Van Laerhoven *et al.* 2008a] Kristof Van Laerhoven, Marko Borazio, David Kilian, and Bernt Schiele. Sustained Logging and Discrimination of Sleep Postures with Low-Level, Wrist-Worn Sensors. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 69–76, Pittsburgh, PA, USA, September 2008. *cited on pp.* 11

[Van Laerhoven *et al.* 2008b] Kristof Van Laerhoven, David Kilian, and Bernt Schiele. Using Rhythm Awareness in Long-Term Activity Recognition. In *Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC'08)*, pages 63–66, Pittsburgh, PA, USA, October 2008. *cited on pp.* 10, 14, 16

[Vapnik 1998] Vladimir Vapnik. *Statistical learning theory*. Wiley and Sons, NY, 1998. *cited on pp.* 27

[Vurgun *et al.* 2007] Sengul Vurgun, Matthai Philipose, and Misha Pavel. A Statistical Reasoning System for Medication Prompting. In *Proceedings of the 9th International Conference on Ubiquitous Computing (UbiComp'07)*, pages 1–18, Innsbruck, Austria, September 2007. *cited on pp.* 11

[Wan 1999] Dadong Wan. Magic Medicine Cabinet: A Situated Portal for Consumer Healthcare. In *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing (HUC'99)*, pages 352–355, Karlsruhe, Germany, September 1999. *cited on pp.* 10

[Wang *et al.* 2007] Shiaokai Wang, William Pentney, Ana-Maria Popescu, Tanzeem Choudhury, and Matthai Philipose. Common Sense Based Joint Training of Human Activity Recognizers. In *Proceeding of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 2237–2242, Hyderabad, India, January 2007. *cited on pp.* 13, 17

[Ward *et al.* 2005] Jamie A. Ward, Paul Lukowicz, and Gerhard Tröster. Gesture Spotting Using Wrist Worn Microphone and 3-Axis Accelerometer. In *Proceedings of the Joint Conference on Smart Objects and Ambient Intelligence (sOc-EUSAI)*, pages 99–104, Grenoble, France, October 2005. *cited on pp.* 10, 13, 16

[Ward *et al.* 2006] Jamie A. Ward, Paul Lukowicz, and Gerhard Tröster. Evaluating Performance in Continuous Context Recognition Using Event-Driven Error Characterization. In *Proceedings of the 2nd International Workshop on Location- and Context-Awareness (LoCA'06)*, pages 239–255, Dublin, Ireland, May 2006. *cited on pp.* 5

[Weiser 1991] Mark Weiser. The Computer for the 21st Century. *Scientific American*, 265(3):94–104, September 1991. *cited on pp.* 1, 9

[Westeyn *et al.* 2005] Tracy Westeyn, Kristin Vadas, Xuehai Bian, Thad Starner, and Gregory D. Abowd. Recognizing Mimicked Autistic Self Stimulatory Behaviors Using HMMs. In *Proceedings of the 9th IEEE International Symposium on Wearable Computers (ISWC'05)*, pages 164–167, Osaka, Japan, October 2005. *cited on pp.* 11

[Westeyn *et al.* 2006] Tracy Westeyn, Peter Presti, and Thad Starner. ActionGSR: A Combination Galvanic Skin Response-Accelerometer for Physiological Measurements in Active Environments. In *Proceedings of the 10th International Symposium on Wearable Computers (ISWC'06)*, pages 129–130, Montreux, Switzerland, October 2006. *cited on pp.* 13

[Westeyn *et al.* 2008] Tracy Westeyn, Julie A. Kientz, Thad Starner, and Gregory Abowd. Designing Toys with Automatic Play Characterization for Supporting the Assessment of a Child's Development. In *Proceedings of the Workshop on Designing for*

children with Special Needs at the 7th International Conference on Interaction Design and Children (IDC'08), pages 89–92, Chicago, IL, USA, June 2008. *cited on pp.* 11

[Wheeler and Reis 1991] Ladd Wheeler and Harry T. Reis. Self-Recording of Everyday Life Events: Origins, Types, and Uses. *Journal of Personality*, 59(3):339–354, 1991. *cited on pp.* 14

[Wilson and Philipose 2005] Daniel H. Wilson and Matthai Philipose. Maximum a Posteriori Path Estimation with Input Trace Perturbation: Algorithms and Applications to Credible Rating of Human Routines. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'05)*, 2005. *cited on pp.* 11

[Wilson *et al.* 2003] Daniel H. Wilson, Danny Wyatt, and Matthai Philipose. Using Context History Data for Data Collection in the Home. In *Proceedings of the Pervasive'05 International Workshop on Exploiting Context Histories in Smart Environments (ECHISE'05)*, Munich, Germany, 2003. *cited on pp.* 14

[Wilson *et al.* 2005] Daniel H. Wilson, Sunny Consolvo, Ken Fishkin, and Matthai Philipose. In-Home Assessment of the Activities of Daily Living of the Elderly. In *Extended Abstracts of CHI 2005: Workshop - HCI Challenges in Health Assessment*, page 2130, Portland, OR, USA, April 2005. *cited on pp.* 2

[Wilson 2005] Daniel H. Wilson. *Assistive Intelligent Environments for Automatic Health Monitoring*. PhD thesis, Carnegie Mellon University, 2005. *cited on pp.* 104

[Witten and Frank 2005] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005. *cited on pp.* 27

[Wren and Tapia 2006] Christopher R. Wren and Emmanuel Munguia Tapia. Toward Scalable Activity Recognition for Sensor Networks. In *Proceedings of the 2nd International Workshop on Location- and Context-Awareness (LoCA'06)*, pages 168–185, Dublin, Ireland, May 2006. *cited on pp.* 12, 22

[Wu *et al.* 2007] Jianxin Wu, Adebola Osuntogun, Tanzeem Choudhury, Matthai Philipose, and James M. Rehg. A Scalable Approach to Activity Recognition based on Object Use. In *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007. *cited on pp.* 13

[Wyatt *et al.* 2005] Danny Wyatt, Matthai Philipose, and Tanzeem Choudhury. Unsupervised Activity Recognition Using Automatically Mined Common Sense. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, pages 21–27, Pittsburg, PA, USA, July 2005. *cited on pp.* 10, 12, 17

[yau Lin and jen Hsu 2006] Chi yau Lin and Yung jen Hsu. IPARS: Intelligent Portable Activity Recognition System via Everyday Objects, Human Movements, and Activity Duration. In *Proceeding of the AAAI Workshop Modeling Others from Observations (MOO 2006)*, Boston, MA, USA, July 2006. *cited on pp.* 12

[Zhou 2004] Zhi-Hua Zhou. Multi-Instance Learning: A Survey. Technical report, AI Lab, Department of Computer Science and Technology, Nanjing University, 2004. *cited on pp.* 65

[Zhu and Ghahramani 2002] Xiaojin Zhu and Zoubin Ghahramani. Learning from Labeled and Unlabeled Data with Label Propagation. Technical report, CMU-CALD-02-107, 2002. *cited on pp.* 84, 86

[Zinnen *et al.* 2007] Andreas Zinnen, Kristof Van Laerhoven, and Bernt Schiele. Toward Recognition of Short and Non-repetitive Activities from Wearable Sensors. In *Proceedings of the European Conference on Ambient Intelligence (AmI'07)*, pages 142–158, Darmstadt, Germany, November 2007. *cited on pp.* 10, 14

[Zinnen *et al.* 2009] Andreas Zinnen, Christian Wojek, and Bernt Schiele. Multi Activity Recognition based on Bodymodel-Dervied Primitives. In *Proceedings of the 4th International Symposium on Location and Context Awareness (LoCA'09)*, Tokyo, Japan, May 2009. *cited on pp.* 103

Curriculum Vitae

Maja Stikic

Date of birth: May 6, 1978 in Sarajevo, Bosnia and Herzegovina

Citizenship: Serbian

Education:

2006–2009 **Technische Universität Darmstadt, Germany**
PhD Studies in Computer Science
Degree: Doktor-Ingenieur (Dr.-Ing.)

1997–2004 **University of Belgrade, Serbia**
Faculty of Electrical Engineering
Studies in Computer Engineering and Information Theory
Degree: Graduate Engineer of Electrical Engineering (Dipl.-Ing.)

1993–1997 **Mathematical Gymnasium, Belgrade, Serbia**
Degree: Secondary School Matriculation

Work Experience:

2007–2009 **Fraunhofer IGD, Darmstadt, Germany**
Doctoral fellowship

2005–2006 **Fraunhofer IPSI, Darmstadt, Germany**
Researcher, doctoral fellowship

2004–2005 **Institute Mihajlo Pupin, Belgrade, Serbia**
Research and Development Engineer

2004 **Fraunhofer IPSI, Darmstadt, Germany**
Internship

Publications

[7] Maja Stikic, Diane Larlus, and Bernt Schiele. Multi-Graph Based Semi-Supervised Learning for Activity Recognition. International Symposium on Wearable Computers (ISWC'09). Linz, Austria, September 2009.

[6] Maja Stikic and Bernt Schiele. Activity Recognition from Sparsely Labeled Data Using Multi-Instance Learning. International Symposium on Location and Context Awareness (LoCA'09), Tokyo, Japan, May 2009.

[5] Maja Stikic, Kristof Van Laerhoven, and Bernt Schiele. Exploring Semi-Supervised and Active Learning for Activity Recognition. IEEE International Symposium on Wearable Computers (ISWC'08), Pittsburgh, PA, USA, September 2008.

[4] Maja Stikic, Tâm Huỳnh, Kristof Van Laerhoven, and Bernt Schiele. ADL Recognition Based on the Combination of RFID and Accelerometer Sensing. International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health 2008), Tampere, Finland, January 2008.

[3] Maja Stikic and Kristof Van Laerhoven. Recording Housekeeping Activities with Situated Tags and Wrist-Worn Sensors: Experiment Setup and Issues Encountered. International Workshop on Wireless Sensor Networks for Health Care (WSNHC'07), held in conjunction with the International Conference of Networked Sensing Systems (INSS'07), Braunschweig, Germany, June 2007.

[2] Fano Ramparany, Remco Poortinga, Maja Stikic, Jörg Schmalenströer, and Thorsten Prante. An open Context Information Management Infrastructure - the IST-Amigo Project. IET International Conference on Intelligent environments (IE'07), Ulm, Germany, September 2007.

[1] Carsten Magerkurth, Timo Engelke, and Maja Stikic. Gesture Based Interaction Techniques in Hybrid Environments. International Conference on Digital Interactive Media Entertainment and Arts (DIMEA '06), Bangkok, Thailand, October 2006.